Author:
Wiffen, Fred D

Title:
Advanced MIMO Techniques for Future Wireless Communications

Advanced MIMO Techniques for Future Wireless Communications

Alfred Daniel Wiffen

Department of Electrical and Electronic Engineering

University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Engineering.

November 2020

Word count: sixty-eight thousand

Advanced MIMO Techniques for Future Wireless Communications

Multi-user multiple input multiple output (MU-MIMO) technologies exploit the spatial domain to serve multiple users on the same time-frequency resource, achieving unrivalled spectral efficiencies. This thesis investigates and proposes novel signal processing-based solutions to practical challenges associated with two MU-MIMO technologies that are expected to play a key role in future wireless systems – massive MIMO and distributed MIMO cloud radio access networks (C-RAN).

The first part of this thesis addresses the problem of peak-to-average-power ratio (PAPR) reduction, for improving the operating power efficiency of the large number of power amplifiers used in massive MIMO transmitters. It begins by using Bussgang's theory to derive a statistical signal model for the distortion introduced by conventional clipping-based PAPR reduction. This model is then used to develop a practical and effective PAPR reduction scheme that uses spatial filtering to eliminate the effects of clipping distortion from the signals received by the users, and can incorporate active constellation extension for improved performance.

The remainder of the thesis focuses on lossy data compression for MIMO C-RAN – reducing the quantity of signal data such that low capacity fronthaul connections can be used. For the massive MIMO uplink, transform coding is shown to be effective at exploiting the inherent sparsity in the received signals to achieve efficient data compression. This transform coding approach is then adapted for distributed MIMO, using jointly optimised rate allocations to account for correlations between the signals received at different remote receivers.

The final part of the thesis shows that distributed dimension reduction can be applied to distributed MIMO to produce a reduced dimension MIMO system that preserves many of the benefits provided by deploying a large number of antennas. Combined with simple scalar quantization, this represents an efficient fronthaul data compression/reduction strategy for both the distributed MIMO uplink and downlink.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: Alfred Daniel Wiffen DATE: 01/11/2020

Acknowledgements

I would first of all like to thank my academic supervisors, Professors Angela Doufexi, Mark Beach and Andrew Nix, for their continued support and advice – often required at short notice! I would like to thank my industrial sponsor, Toshiba BRIL, for the financial support they have provided both during the PhD and since. My particular thanks go to my supervisor at Toshiba, Dr Zubeir Bocus, for the help he gave me in finding a direction for this research, and Dr Woon Hau Chin, who has supported me since Zubeir's departure.

Finally, I would like to thank my family, without whose support I may not have completed this PhD; my friends in Bristol and beyond, without whose company I may have completed it a fair bit sooner; and my colleagues in CSN, whom without me may have completed theirs a little earlier too.

"We are stuck with technology
when what we really want
is just stuff that works."

DOUGLAS ADAMS

Contents

List of Figures viii

List of Abbreviations xiv

List of Notations xvi

1 Introduction 1
- 1.1 Advanced Multi-User MIMO . . . 1
- 1.2 Research Principles . . . 3
- 1.3 Thesis Structure & Key Contributions . . . 4
- 1.4 Publications . . . 6

2 Fundamentals of Multi-User MIMO Communication 8
- 2.1 Wireless Digital Communication . . . 9
  - 2.1.1 The Wireless Channel . . . 9
  - 2.1.2 Modulation . . . 11
  - 2.1.3 Channel Capacity . . . 14
  - 2.1.4 Fading Channels . . . 17
- 2.2 From SISO to MU-MIMO . . . 22
  - 2.2.1 SIMO & MISO . . . 23
  - 2.2.2 MU-MIMO . . . 26
  - 2.2.3 Channel Capacity . . . 28
- 2.3 MU-MIMO Processing . . . 32
  - 2.3.1 Linear Detection . . . 32
  - 2.3.2 Linear Precoding . . . 36
  - 2.3.3 Non-Linear Methods . . . 38
- 2.4 MU-MIMO Channels . . . 42
  - 2.4.1 Correlated Fading Channel Models . . . 42
  - 2.4.2 Channel Estimation . . . 44
  - 2.4.3 Power Control . . . 49
- 2.5 Advanced MU-MIMO Architectures . . . 52
  - 2.5.1 Massive MIMO . . . 52
  - 2.5.2 Distributed MIMO C-RAN . . . 58
- 2.6 Conclusion . . . 65

3 Clipping-based PAPR Reduction for the Massive MIMO Downlink 66
3.1 Chapter Overview . . . 68
3.1.1 Novel Contributions . . . 68
3.1.2 Published Work . . . 69
3.2 Background . . . 69
3.2.1 Motivation . . . 70
3.2.2 Classical PAPR Reduction . . . 73
3.2.3 PAPR Reduction in MIMO Systems . . . 75
3.3 Clipping & Filtering in Massive MIMO-OFDM Systems . . . 80
3.3.1 Clipping & Filtering of MIMO-OFDM . . . 80
3.3.2 Iterative Clipping & Filtering . . . 83
3.3.3 Phase-only OFDM . . . 84
3.4 Impact of Clipping & Filtering on Massive MIMO Performance . . . 85
3.4.1 Asymptotic Performance . . . 87
3.4.2 Spatially Correlated Channels . . . 88
3.4.3 Near-Far Effect . . . 90
3.5 Clipping & Least Squares Spatial Filtering . . . 91
3.5.1 Least Squares Spatial Filtering / Nullspace Projection Method . . . 92
3.5.2 Bussgang-aware Least Squares Filtering . . . 94
3.5.3 Iterative BLS . . . 96
3.5.4 BLS in Correlated Channels . . . 99
3.5.5 Active Constellation Extension . . . 100
3.5.6 Practical Aspects . . . 104
3.6 Conclusion . . . 106

4 Transform Coding-based Signal Compression for Uplink MIMO C-RAN 109
4.1 Chapter Overview . . . 112
4.1.1 Novel Contributions . . . 112
4.1.2 Published Work . . . 113
4.2 Background . . . 113
4.2.1 Motivation . . . 113
4.2.2 Fundamentals of Signal Compression . . . 116
4.2.3 Signal Compression for Uplink MIMO C-RAN . . . 120
4.3 Massive MIMO Uplink Signal Compression . . . 124
4.3.1 The Limits of Sample & Forward . . . 124
4.3.2 Transform Coding for the Massive MIMO Uplink . . . 127
4.3.3 Transform Coding with a Uniform Quantization Noise Level . . . 132
4.3.4 Massive MIMO with Limited Fronthaul . . . 135
4.3.5 Transform Coding with Imperfect CSI . . . 141
4.4 Distributed MIMO Uplink Signal Compression . . . 144
4.4.1 Transform Coding with a Uniform Quantization Noise Level . . . 145
4.4.2 Optimal Transform Coding for the Distributed MIMO Uplink . . . 147
4.4.3 Transform Coding with Jointly Optimised Rate Allocation . . . 148

4.4.4 Distributed MIMO with Limited Fronthaul . . . 154
4.4.5 Transform Coding with Imperfect CSI . . . 157
4.4.6 Practical Aspects . . . 158
4.5 Conclusion . . . 161

5 Dimension Reduction for Distributed MIMO C-RAN 164
5.1 Chapter Overview . . . 167
5.1.1 Novel Contributions . . . 168
5.1.2 Published Work . . . 169
5.2 Background . . . 169
5.2.1 Distributed Dimension Reduction . . . 170
5.2.2 Distributed Dimension Reduction in MIMO C-RAN . . . 172
5.3 Reduced Dimension Distributed MIMO Uplink Channels . . . 173
5.3.1 Maximum Mutual Information Distributed Dimension Reduction . . . 174
5.3.2 Reduced Dimension Channels . . . 178
5.3.3 Matched Filter-based Distributed Dimension Reduction . . . 182
5.3.4 Dimension Reduction with Imperfect CSI . . . 188
5.4 Reduced Dimension MIMO Uplink with Lossy Compression . . . 190
5.4.1 Sum Capacity under Gaussian Scalar Compression . . . 190
5.4.2 Practical Reduced Dimension Compression using Fixed-Rate Quantizers . 196
5.4.3 Example: Dense Deployment . . . 200
5.5 Two-Stage Reduced Dimension Precoding for the Distributed MIMO Downlink . 202
5.5.1 Two-Stage Precoder Design . . . 203
5.5.2 Quantization-aware Power Allocation . . . 205
5.5.3 Example: Dense Deployment . . . 207
5.6 Conclusion . . . 208

6 Conclusion 211

Appendices 217
1 Clipping-based PAPR Reduction for the Massive MIMO Downlink . . . 217
1.1 Least Squares Derivation . . . 217
1.2 Optimal Target Symbol Scaling . . . 218
2 Transform Coding-based Signal Compression for Uplink MIMO C-RAN . . . 218
2.1 Upper Bound on Sample & Forward Capacity . . . 218
2.2 SCA - Joint Rate Allocation . . . 219
3 Dimension Reduction for Distributed MIMO C-RAN . . . 219
3.1 Conditional Entropy . . . 219
3.2 Determinant Maximisation . . . 220

Bibliography 221

List of Figures

1.1 Space division multiple access using massive MIMO. 2
1.2 Space division multiple access using distributed MIMO C-RAN. 3

2.1 Basic wireless communication system block diagram. 9
2.2 Equivalent digital channel. 10
2.3 Channel impulse response (top) and associated OFDM subcarrier gains with $L = 12$, $N = 128$. 13
2.4 Bit error rate performance of QPSK with 3/5 rate LDPC coding. 15
2.5 AWGN channel capacity, $|h|^2 = 1$. 16
2.6 Wireless channel with direct and two reflected paths. 18
2.7 Outage probability for Rayleigh channel with different diversity levels, $\mathcal{C}^\star = 1$ bpcu. 21
2.8 SISO channel capacity. 22
2.9 SIMO system with $M = 3$. 23
2.10 CDF of SIMO channel capacity, $\rho = 10$ dB 25
2.11 MISO system with $M = 3$. 26
2.12 MU-MIMO system with $M = 3$ and $K = 2$. 27
2.13 MU-MIMO sum capacity under i.i.d Rayleigh fading, $\rho_k = P/K$, $M = 8$ 30
2.14 MIMO channel condition under i.i.d Rayleigh fading, $M = 8$. 31
2.15 MIMO capacity under linear detection with i.i.d Rayleigh fading, top: $M = 8, K = 4$, bottom: $M = 8, K = 8$ 36
2.16 Bit error rate of QPSK modulation with MMSE, MMSE V-BLAST (random & optimised cancellation orders) and sphere decoding in i.i.d Rayleigh channels, top: $M = 8, K = 2$, bottom: $M = 8, K = 6$ 41
2.17 Two planewaves incident on four-element 1-dimensional array. 43
2.18 Uplink channel capacity with varying quality of CSI, $M = 8, K = 4$. Solid lines: optimal detection, dotted lines: MMSE detection. 49
2.19 Distribution of downlink user capacities under average and max-min power control, ZF beamforming, $M = 8, K = 4, P_T = 10$ dB, i.i.d Rayleigh fading. 51
2.20 Intensity plots of $\frac{1}{M}\mathbf{H}^\dagger\mathbf{H}$ for various values of M, i.i.d Rayleigh fading channel, $K = 8$. 54
2.21 Mean sum capacity under linear detection with varying M, i.i.d Rayleigh fading, $K = 10$, $p_k\beta_k = 1$. Top: $\gamma = -5$ dB, bottom: $\gamma = 15$ dB. 55

2.22 CDF of user capacities under MMSE detection with varying M, i.i.d Rayleigh fading, $\gamma = 10$ dB, $p_k\beta_k = 1$, $K = 10$. 56
2.23 User capacity scaling with SNR for different numbers of antennas, i.i.d Rayleigh fading, $K = 8$, $p_k\beta_k = 1$. 56
2.24 Example distributed MIMO C-RAN configuration with $K = 8, L = 4, M = 8$. . . 60
2.25 User uplink mean capacities for varying number of distributed receivers, L, i.i.d Rayleigh fading, $K = 4$, $ML = 16$. Solid line: mean capacity, dashed line: 10% outage capacity. 62
2.26 User uplink capacities for varying number of antennas per receiver, i.i.d Rayleigh fading, $K = 8$, $L = 4$. Solid line: mean capacity, dashed line: 10% outage capacity. 62
2.27 Mean downlink user capacity under ZF precoding with max-min power control, i.i.d Rayleigh fading, $K = 8$, $L = 4$. Solid line: standard ZF, dashed line: ZF-MPC. 64

3.1 Block diagram of proposed PAPR reduction scheme. 68
3.2 Example power amplifier input-output characteristic. 71
3.3 Complementary cumulative distribution function of OFDM PAPR (baseband), varying number of subcarriers (N), QPSK symbols. 72
3.4 Active constellation extension regions (shaded) for QPSK. 74
3.5 Complementary cumulative distribution function of MIMO-OFDM PAPR, varying number of BS antennas (M), 10 time domain i.i.d Rayleigh fading channel taps, 512 subcarriers, QPSK symbols. 76
3.6 Complementary cumulative distribution function of MIMO-OFDM PAPR in correlated channel, with different precodings. 77
3.7 Original OFDM signal (top), clipped OFDM signal (middle), clipped & filtered OFDM signal (bottom). 83
3.8 Complementary cumulative distribution function of array PAPR with iterative clipping & filtering, $\gamma = 1.2$, 8 users, 64 BS antennas, 512 subcarriers, QPSK symbols. 84
3.9 Mean user capacity for different clipping ratios with and without power normalisation, clipping & filtering with three iterations. Normalised to receive SNR 15 dB (unclipped), 8 users, 64 BS antennas, 512 subcarriers. 86
3.10 Mean user capacity with varying antenna numbers, 8 users, 512 subcarriers, $\gamma = 1.2$, 3 iterations. 88
3.11 Solid line: normalised eigenvalues of clipping noise covariance (measured on central subcarrier), dashed line: normalised eigenvalues of transmit signal covariance. 89
3.12 Mean user capacity with varying antenna numbers, 8 users, 512 subcarriers, $\gamma = 1.2$, 3 iterations. 90
3.13 Near-far effect of clipping on user capacity where average channel strength of near user is 10 dB greater than far users. Power normalised to receive SNR 15 dB (unclipped), three iterations of clipping & filtering, 1 near user, 7 far users, 64 BS antennas, 512 subcarriers. 91
3.14 PAPR reduction of iterative least squares filtering scheme in Rayleigh channel, $K = 8$, $M = 64$ and $N = 512$. 94

3.15 Comparison of PAPR regrowth of LS and BLS spatial filtering methods. 96
3.16 PAPR reduction of iterative BLS scheme in Rayleigh channel, $K = 8$, $M = 64$ and $N = 512$. 97
3.17 Loss in received signal power after power normaliation. 5 clipping iterations, i.i.d Rayleigh channel, $K = 8$, $M = 64$ and $N = 512$. 98
3.18 PAPR reduction and receive SNR loss of iterative BLS clipping & filtering in 10 tap i.i.d Rayleigh channel, $K = 8$, $M = 64$ and $N = 512$. 98
3.19 Bit error rate in near-far scenario, i.i.d Rayleigh channel, QPSK, 5 clipping iterations with $\gamma = 1$, $K = 8$, $M = 64$ and $N = 512$. 99
3.20 PAPR reduction and receive SNR loss of iterative BLS clipping & filtering in correlated Rayleigh channel, $K = 8$, $M = 64$ and $N = 512$. 100
3.21 Illustration of projection of $\mathbf{h}_k^T \mathbf{x}_{\text{CF}}$ into $C(\mathbf{s})$ for QPSK. 101
3.22 PAPR of ACE PAPR reduction scheme in 10 tap i.i.d Rayleigh channel. 5 clipping iterations, $\gamma = 1$, $K = 8$, $M = 64$, $N = 512$. 102
3.23 PAPR of ACE PAPR reduction scheme in 10 tap i.i.d Rayleigh channel. 5 clipping iterations, $\gamma = 1$, $K = 8$, $M = 16$, $N = 512$. 102
3.24 BER of QPSK under CF, BLS and ACE PAPR reduction schemes in 10 tap i.i.d Rayleigh channel. 5 clipping iterations, $\gamma = 1$, $K = 8$, $M = 16$, $N = 512$. 103

4.1 Block diagram of proposed transform coding fronthaul compression scheme. . . . 110
4.2 Fronthaul constrained C-RAN network topology with $K = 5$ users, $L = 2$ receivers and central processor. 115
4.3 Exact and approximate quantization noise variance. 117
4.4 Cut-set upper bound on sum capacity of C-RAN network with limited fronthaul capacity. 120
4.5 Single cell C-RAN uplink network topology 124
4.6 Sum capacity upper bound under sample & forward for different antenna ratios, i.i.d Rayleigh fading, $K = 8$. 126
4.7 Sum capacity under sample & forward for different SNRs, i.i.d Rayleigh fading, $K = 8$, $M = 32$. 127
4.8 Normalised sum capacity of massive MIMO system under transform coding with varying SNR, i.i.d Rayleigh channel, $K = 8$, $M = 64$. 130
4.9 Comparison of sum capacity of massive MIMO with transform coding and sample & forward compression with varying number of antennas, M, i.i.d Rayleigh channel, $K = 8$, $\rho = 10$ dB. 132
4.10 Scalar quantization noise variances under waterfilling and UQN rate allocations, i.i.d Rayleigh fading channel, $K = 8$, $M = 64$, $\rho = 10$ dB. Coloured lines show the quantization noises on different eigenchannels. 134
4.11 Sum capacity under transform coding with KLT transform and waterfilling & UQN rate allocations, i.i.d Rayleigh fading channel, $K = 8$, $M = 64$. 135
4.12 Mean and outage user capacities with varying number of BS antennas, M, fixed fronthaul capacity $\mathcal{R} = 24$ bpcu, i.i.d Rayleigh fading, $K = 8$. 136

4.13 Sum capacity under non-linear and linear detection for different numbers of antennas, i.i.d Rayleigh fading, $K = 8$. Top figure: $\rho = 10$ dB, bottom figure: $\rho = -10$ dB. 137
4.14 Sum capacity scaling with SNR for different numbers of antenna under a fixed fronthaul capacity, i.i.d Rayleigh fading, $K = 8$. Top figure: $\mathcal{R} = 32$ bpcu, bottom figure: $\mathcal{R} = 64$ bpcu. 138
4.15 User capacity CDF for different numbers of antennas under a fixed fronthaul capacity $\mathcal{R} = 24$ bpcu, i.i.d Rayleigh fading channel, $K = 8$. 139
4.16 Sum capacity under non-linear and linear detection for different numbers of antennas, correlated Rayleigh fading, $\rho = 10$ dB, $K = 8$. 140
4.17 Mean and outage user capacities with varying number of BS antennas, M, fixed fronthaul capacity $\mathcal{R} = 24$ bpcu, correlated Rayleigh fading, $K = 8$. 140
4.18 Sum capacity scaling with varying estimated CSI quality, i.i.d Rayleigh fading channel, $\rho = 10$ dB, $K = 8$, $M = 64$. 143
4.19 Sum capacity scaling with varying SNR, ρ, i.i.d Rayleigh fading channel, $\rho_{\text{CSI}} = 15$ dB, $K = 8$, $M = 64$. 144
4.20 Distributed MIMO sum capacity under UQN compression, $\rho = 10$ dB, i.i.d Rayleigh fading channel, $K = 8$, $L = 4$ $M = 16$. 146
4.21 Distributed MIMO sum capacity with optimal point-to-point compression, $\rho = 10$ dB, i.i.d Rayleigh fading channel, $K = 8$, $L = 4$ $M = 8$. 148
4.22 Distributed MIMO sum capacity with KLT and optimal SCA rate allocation, i.i.d Rayleigh fading channel, $K = 8$, $L = 8$ $M = 4$, $\rho = 0$ dB. 152
4.23 Distributed MIMO sum capacity with KLT and optimal SCA rate allocation, i.i.d Rayleigh fading channel, $\rho = 10$ dB, $K = 8$, $L = 4$ $M = 8$. 153
4.24 Average (mode) number of signal components allocated $r_{l,i} \geq 0.1$ bpcu at different fronthaul capacities, i.i.d Rayleigh fading channel, $K = 8$, $M = 8$, $L = 4$. . . 154
4.25 Fronthaul utilisation, $K = 8$, $M = 8$, $L = 4$. 155
4.26 Mean user throughputs in 100 MHz channel for different user & RRH densities, i.i.d Rayleigh fading channel, $M = 8$. Solid line: $K = 8, L = 4, \rho = 10$ dB, dot-dash line: $K = 16, L = 8$, dashed line: $K = 32, L = 16$. 156
4.27 Distributed MIMO mean user throughputs in 100 MHz channel with varying numbers of receiver antennas, i.i.d Rayleigh fading channel, $K = 8, L = 4$. Top: per-receiver fronthaul throughput 1.5 Gbps, bottom: per-receiver fronthaul throughput 3 Gbps. 157
4.28 SCA-RA with imperfect CSI, i.i.d Rayleigh fading channel, $K = 8$, $L = 4$ $M = 8$, $\rho = 10$ dB. 158
4.29 SCA-RA with limited number of iterations, $\rho = 10$ dB, i.i.d Rayleigh fading channel, $K = 8$, $L = 4$ $M = 8$. 159

5.1 Block diagram of proposed dimension reduction uplink fronthaul compression scheme (single receiver shown). 165
5.2 Block diagram of proposed two-stage fronthaul-aware downlink precoding scheme. 166

5.3 Convergence of T-CKLT BCA algorithm, $\rho = 15$ dB, $M = 8$. Solid line: $K = 8, L = 4$, dashed line: $K = 16, L = 8$. 177
5.4 Reduced dimension mutual information scaling with ρ, $K = 8$, $M = 8$, $L = 4$ ($t = 8$). Solid line: T-CKLT filters, dot-dash line: random semi-orthogonal filter. 179
5.5 Reduced dimension channel eigenvalue spread, $K = 8$, $M = 8$, $L = 4$ ($t = 8$). Top: T-CKLT filters, bottom: random semi-orthogonal filters. 180
5.6 Array gain under dimension reduction, $N = 3$, $K = 8$, $L = 4$. Solid line: T-CKLT, dot-dash line: random filtering (all M). 181
5.7 Channel hardening under dimension reduction, $M = 8$, $K = 8$, $L = 4$. Solid line: T-CKLT, $N = 4$, dashed line: T-CKLT, $N = 3$, dot-dash line: random filtering, $N = 3$ (all values of M). 181
5.8 Performance of different dimension reduction schemes, $K = 8$, $L = 4$. Solid line: T-CKLT, dot-dash line: MF-GS, dotted line: antenna selection. 185
5.9 Channel eigenvalue spread under dimension reduction, $K = 8$, $M = 8$, $L = 4$. Solid line: MF-GS, dashed line: F-MF-GS, $N_s = 4$. 187
5.10 Illustration of reduced CSI signalling overheads facilitated by F-MF-GS dimension reduction scheme. 187
5.11 Average signalling overheads associated with different dimension reduction schemes, channel coherence block size 100 symbols, $K = 8$, $M = 8$, $n_b = 20$ bits. 188
5.12 Joint mutual information lower bound under T-CKLT dimension reduction, $K = 8, M = 8, L = 4$. 189
5.13 Sum capacity under RQ-BCA dimension reduction filters for different signal dimensions, $\rho = 15$ dB, $K = 8, M = 8, L = 4$ (single channel realisation). 192
5.14 Asymptotic high SNR sum capacity scaling under T-CKLT dimension reduction, $K = 8, M = 8, L = 4$. Blue lines: $N = 2$, orange lines: $N = 4$. 194
5.15 Sum capacity scaling under T-CKLT dimension reduction, $\rho = 15$ dB,$K = 8, M = 8, L = 4$. Blue lines: $N = 2$, orange lines: $N = 3$, yellow lines: $N = 4$. . . 194
5.16 Sum capacity performance under different dimension reduction schemes, $\rho = 15$ dB. Top: $K = 8, M = 8, L = 4$, bottom: $K = 8, M = 4, L = 8$. 196
5.17 User symbol mean squared error after MMSE detection, T-CKLT dimension reduction, fixed-rate scalar quantizers, $K = 8, M = 8, L = 4$. Solid line: analytical expression, dashed line: Gaussian symbols, dot-dash line: 64QAM symbols. . . . 198
5.18 User mean and 10 % outage capacities under fixed-rate scalar quantization with varying signal dimensions and quantizer resolutions, $\rho = 15$ dB, $K = 8, M = 8, L = 4$. Top: T-CKLT dimension reduction filters, bottom: MF-GS dimension reduction filters. 199
5.19 User mean and outage capacities under fixed-rate scalar quantization with MF-GS dimension reduction filters, $K = 8, L = 4, N = 3, b = 10$ bits ($\mathcal{R} = 30$ bpcu). 200
5.20 Mean user throughputs in dense distributed MIMO deployment with MF-GS dimension reduction and different channel estimation qualities, $N = 2$, $b = 8$ bits, $K = 16, L = 16, M = 4$. 201

5.21 Mean user throughputs in dense distributed MIMO deployment with MF-GS dimension reduction, $\rho_{\text{CSI}} = 20$ dB, $N = 2, K = 16, L = 16$. Solid line: per-receiver fronthaul throughput 960 Mbps, dot-dash line: fronthaul throughput 1200 Mbps, dashed line: fronthaul throughput 1440 Mbps. 202
5.22 Mean downlink user capacities under two-stage precoding with T-CKLT outer precoder design and ZF/ZF-MPC inner precoder design. $P_T = 24$ dBm, $B = 20$ MHz, receiver noise figure 5 dB, $K = 8, L = 4, M = 8$. 205
5.23 Performance of two-stage precoding with and without quantization-aware max-min power control, MF-GS outer precoder design, $P_T = 24$ dBm, $B = 20$ MHz, noise figure 5 dB, $K = 8, L = 4, M = 8, N = 3$. Top: standard ZF inner precoder, bottom: ZF-MPC inner precoder design. Blue line: $b = 8$ bpcu, orange line: $b = 10$ bpcu, yellow line: $b = 12$ bpcu. 207
5.24 Mean user downlink throughputs in dense distributed MIMO deployment with MF-GS outer precoders, $B = 60$ MHz, receiver noise figure 5 dB, $N = 2, K = 16, L = 16$. Solid line: per-transmitter fronthaul throughput 960 Mbps, dot-dash line: fronthaul throughput 1200 Mbps, dashed line: fronthaul throughput 1440 Mbps. 208

List of Abbreviations

ACE	Active constellation extension
ADC	Analogue-to-digital convertor
AWGN	Additive white Gaussian noise
BCA	Block coordinate ascent
BER	Bit error rate
BLS	Bussgang-aware least squares
BS	Basestation
C-RAN	Cloud radio access network
CDF	Cumulative distribution function
CE	Constant envelope
CoMP	Co-ordinated multipoint
CP	Central processor
CS	Compressed sensing
CSI	Channel state information
DAC	Digital-to-analogue convertor
DL	Downlink
DPC	Dirty paper coding
F-MF-GS	Fading matched filter Gram-Schmidt
FDD	Frequency division duplex
FFT	Fast Fourier transform
IFFT	Inverse fast Fourier transform
IQ	In-phase & quadrature
ISI	Inter-symbol interference
KLT	Karhunen-Loeve transform
LDPC	Low density parity check
LMMSE	Linear minimum mean squared error
LNA	Low noise amplifier
LoS	Line of sight
LS	Least squares
LTE	Long term evolution
LTE-A	Long term evolution advanced
LTI	Linear time invariant
MAC	Medium access control
MF	Matched filter
MF-GS	Matched filter Gram-Schmidt
MIMO	Multiple input multiple output
MISO	Multiple input single output
ML	Maximum likelihood
MMSE	Minimum mean squared error
MMSE-SIC	Minimum mean squared error with successive interference cancellation
MRC	Maximum ratio combining
MRT	Maximum ratio transmission
MU-MIMO	Multi-user MIMO

OFDM	Orthogonal frequency division multiplexing
OFDMA	Orthogonal frequency division multiple access
PA	Power amplifier
PAPC	Per-antenna power constraint
PAPR	Peak-to-average power ratio
PC	Power control
PF	Phase-only and filtering
PHY	Physical layer
QAM	Quadrature amplitude modulation
QPSK	Quadrature phase shift keying
RF	Radio frequency
RIP	Restricted isometry property
RQ-BCA	Rayleigh quotient block coordinate ascent
RRH	Remote radio head
SCA	Successive convex approximation
SCA-P2P	Successive convex approximation point-to-point
SCA-RA	Successive convex approximation rate allocation
SDMA	Space division multiple access
SIMO	Single input multiple output
SINR	Signal-to-interference-plus-noise ratio
SISO	Single input single output
SNR	Signal-to-noise ratio
SQNR	Signal-to-quantization-plus-noise ratio
SSE	Sum spectral efficiency
T-CKLT	Truncated conditional Karhunen-Loeve transform
TDD	Time division duplex
TPC	Total power constraint
UL	Uplink
ULA	Uniform linear array
UQN	Uniform quantization noise
V-BLAST	Vertical-Bell Laboratories layered space-time
ZF	Zero forcing
ZF-MPC	Zero forcing precoding under multiple power constraints
ZF-TPC	Zero forcing precoding under total power constraint

List of Notations

$\mathbb{R}$	set of real numbers
$\mathbb{C}$	set of complex numbers
$\mathbb{Z}$	set of integers
a	scalar
$\Re(a)$	real part
$\Im(a)$	imaginary part
$\lvert a \rvert$	absolute value
$\mathbb{E}[a]$	expectation
$\mathbb{P}\mathrm{r}[x_i]$	probability of event x_i
$\min(a, b)$	minimum of a & b
$a \gg b$	a much greater than b
$\mathbf{a}$	column vector
$\mathbf{A}$	matrix
$\mathbf{A}^*$	conjugate
$\mathbf{A}^T$	transpose
$\mathbf{A}^\dagger$	conjugate transpose
$\mathrm{Tr}\left(\mathbf{A}\right)$	trace
$\det\left(\mathbf{A}\right)$	determinant
$\left\lVert\mathbf{A}\right\rVert^2$	2-norm, $\mathrm{Tr}\left(\mathbf{A}^\dagger\mathbf{A}\right)$
$\left[\mathbf{A}\right]_{i,j}$	element i, j
$\mathbf{A} = \mathrm{diag}\left(a_1, \ldots, a_N\right)$	$\left[\mathbf{A}\right]_{i,i} = a_i$
$\mathbf{A} > \mathbf{B}$	elementwise matrix inequality
$\mathbf{I}_N$	identity matrix dimension N
$\lvert\mathcal{S}\rvert$	size of set $\mathcal{S}$

Chapter 1

Introduction

Just over a decade on from the release of the first smartphones in 2008, adoption in the UK stands at over 80%, with the average user spending 2.5 hours a day using a smartphone [159] and consuming 3.5 gigabytes of mobile data per month [160]. Over this period, global mobile data traffic is estimated to have grown by a factor of 1000, and now accounts for over 10% of all internet traffic [57], [58]. This growth in connectivity is not limited to Europe or North America – the total number of mobile cellular subscriptions now exceeds the number of people on the planet [137], with usage growing in all regions of the world. In the coming years demand for mobile data is expected to continue to grow exponentially, with high definition video accounting for an increasingly dominant proportion of traffic, and billions of additional 'machine-to-machine' connections massively increasing the number of connected devices [58].

In order to meet this increasing demand for mobile data services & connectivity, future cellular & wireless systems must deliver a step change in performance. Along with providing support for a broader range of devices and quality of service requirements, upcoming fifth generation networks are expected to provide 10, 20 and 100 fold increases in ubiquitous data rate, peak data rate and area capacity, respectively, compared to fourth generation LTE networks [189]. Whilst previous cellular systems have relied heavily on the use of increased transmission bandwidth to improve data rates and capacity, the relative scarcity of high quality sub-6 GHz spectrum available to fifth generation systems means that achieving these targets will require spectral resources to be utilised much more efficiently. Furthermore, with growing concerns around the energy consumption and environmental impact of cellular networks [83], fifth generation systems will be required to simultaneously deliver a 100 fold improvement in energy efficiency [189].

1.1 Advanced Multi-User MIMO

In conventional cellular systems, base stations (BS) are broadly limited to serving a single user within a given frequency band at any particular time instant, to avoid interference between user transmissions. Multiple users are served by dividing up the available time and frequency transmission slots, and the cell capacity and spectral efficiency are fundamentally limited by the available transmit power, propagation conditions and interference from adjacent cells.

Multi-user MIMO (multiple input multiple output) is a physical layer technology that helps

overcome these limitations by using an array of BS antennas and dedicated signal processing to separate user transmissions in the spatial domain, enabling data transmission to or from multiple users to be *spatially multiplexed* within the same time-frequency resource with minimal inter-user interference. Using this *space division multiple access* (SDMA), cell sum capacities and spectral efficiencies that scale linearly with the number of user data streams can be achieved [200].

The core theory behind MU-MIMO has long been well understood [64], but, despite its inclusion in previous-generation LTE standards, a range of practical challenges have meant that the potential benefits of MU-MIMO have yet to be fully realised in commercial systems [117].

The past 10 years have seen significant developments in MU-MIMO technology. Research demonstrating the benefits of deploying a very large number of antennas at the BS has generated a huge amount of interest in so-called *massive* MIMO, ushering in a new MIMO paradigm [178]. These massive MIMO systems use the high spatial resolution provided by a large array of antennas to tightly focus their radiation within the propagation environment, improving performance and radiated energy efficiency whilst simultaneously providing a scalable architecture for achieving the step change in capacity and spectral efficiency required by fifth generation services.

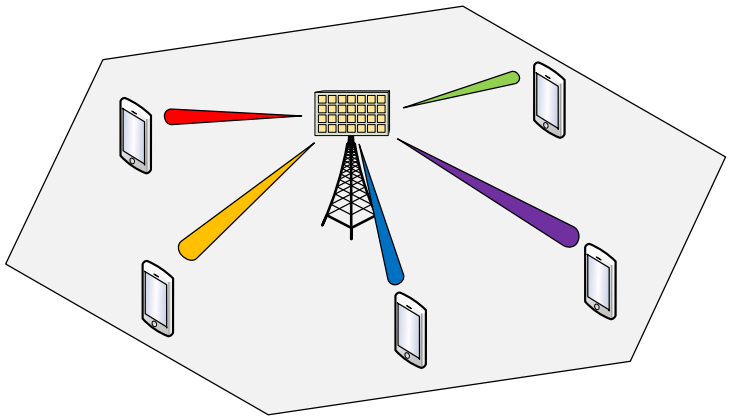

Figure 1.1: Space division multiple access using massive MIMO.

Massive MIMO is now considered a key enabling technology for fifth generation and future wireless systems [5]. However, along with many benefits and opportunities it also brings new challenges, many related to the cost, complexity and power consumption of the hardware associated with deploying a large antenna array [113]. The research in Chapter 3 of this thesis addresses one of these challenges: reducing the peak-to-average power ratio (PAPR) of the precoded orthogonal frequency division multiplexing (OFDM) signals used in the massive MIMO downlink, in order to reduce hardware cost & improve overall energy efficiency.

The increasingly popular cloud radio access network (C-RAN) architecture promises to unlock further enhancements in MU-MIMO technology. In this architecture, the signal processing for multiple different antenna units, or remote radio heads, is performed centrally at a shared central processor (CP), reducing the cost of network deployment and operation [2]. It also opens the door for joint processing of the uplink and downlink signals for multiple remote radio

heads, and a *distributed* MIMO system in which users are jointly served by multiple groups of antennas distributed geographically across the service area – improving energy efficiency by reducing propagation distances whilst providing resilience against signal blockage [86].

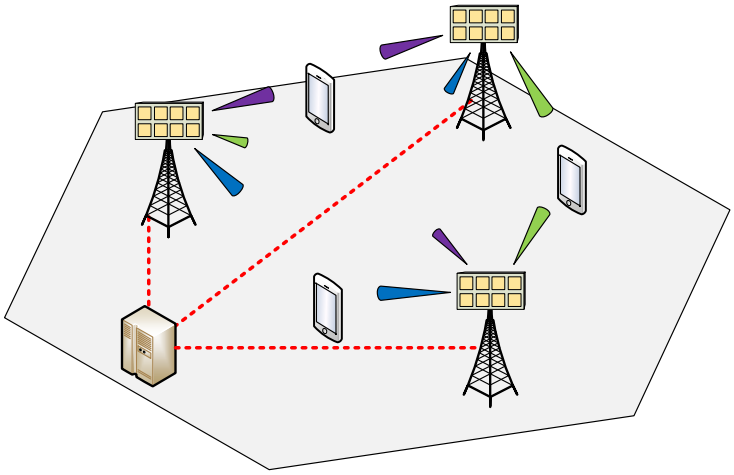

Figure 1.2: Space division multiple access using distributed MIMO C-RAN.

In the C-RAN architecture *fronthaul* connections are used to transfer data between the CP and remote units. Whilst in some scenarios dedicated fibre connections may be used, deployment costs mean that reduced capacity alternatives based on ethernet, shared fibre or wireless point-to-point links are often preferred [98]. In these scenarios, the capacity of the fronthaul network can limit the capacity and spectral efficiency of the distributed MIMO system. The research in Chapters 4 & 5 of this thesis investigates two different data compression strategies for reducing the amount of fronthaul signalling data and maximising MIMO C-RAN performance.

1.2 Research Principles

The research outlined in this thesis uses a combination of mathematical analysis and numerical simulations to develop practical signal processing-based solutions to the massive MIMO PAPR reduction and MIMO C-RAN fronthaul data compression problems. Throughout this work an emphasis is placed on identifying robust, low complexity solutions that could feasibly be implemented in practical systems.

Within the MIMO research literature the use of numerical optimisation techniques is extremely common, and can provide valuable insights & benchmarks. However, many of these techniques are impractical (or impossible) to implement within the constraints of real time operation. A conscious effort was therefore made to seek solutions that use linear processing and simple closed form expressions where possible.

Development of the proposed solutions is guided by broad assumptions & principles, with care taken to avoid overreliance on assumptions, e.g about the propagation environment, that may not be fulfilled in practice. The early stages of the PhD involved participation in some of the Bristol is Open massive MIMO testbed field trials [81]. Whilst implementing any of the final proposed solutions on the testbed was outside of the scope of this work, witnessing first-hand

some of the limitations of the assumptions & simplifications commonly used in the literature[1] had a significant influence on shaping the approach taken in the PhD.

1.3 Thesis Structure & Key Contributions

This thesis has four main chapters – one background chapter and three research chapters:

Chapter 2 introduces the basic MU-MIMO concept in more detail, outlining the core underlying theory and main signal processing operations. Particular attention is given to the material required by subsequent research chapters: capacity equations, performance metrics, propagation models, and various useful mathematical expressions. The chapter ends by describing the important features of the massive MIMO & distributed MIMO C-RAN architectures, with a brief discussion of the main research challenges.

Chapter 3 addresses the problem of PAPR reduction for the massive MIMO downlink. Large instantaneous power fluctuations are a long-standing issue with the OFDM waveform, and increase the power amplifer peak power requirements whilst reducing its operating power efficiency. The precoded signals used on the massive MIMO downlink exacerbate this problem, and hence effective PAPR reduction is required if high energy efficiency is to be achieved.

This chapter proposes a PAPR reduction scheme that adapts the conventional iterative clipping & filtering scheme to include a novel least squares spatial filtering stage that eliminates the effects of clipping noise from the received signals. The main contributions of Chapter 3 are:

- A statistical model based on Bussgang's theorem that accurately statistically models the signal distortion caused by applying clipping to the precoded MIMO signals.
- A novel clipping & spatial filtering PAPR reduction scheme that incorporates the Bussgang clipping model and achieves 8 dB PAPR reduction whilst incurring less than 0.5 dB link degradation.
- An adaptation of the proposed scheme that includes active constellation extension to achieve an additional 1-2 dB of PAPR reduction when smaller symbol constellations are used.

Section 3.1.1 summarises these contributions in more detail.

Chapter 4 addresses the problem of efficiently applying lossy data compression to the received signals on the MU-MIMO C-RAN uplink. The quantity of sampled data produced by the analogue to digital convertors (ADCs) in a remote receiver with multiple antennas can easily exceed the available fronthaul capacity when connections based on ethernet, shared/low grade fibre or wireless links are used. Lossy data compression must therefore be applied at each receiver before the received signals can be transferred to the CP for processing. However, the

[1]For example, the asymptotic optimality of matched filter-based processing in massive MIMO is one of the most widely re-used results from [133]; seeing its performance in real life provided a valuable lesson on the practical limitations of asymptotic analysis.

rudimentary data compression techniques currently used severely limit the achievable MIMO performance when the capacity of the fronthaul network is limited.

This chapter investigates the use of transform coding for achieving efficient uplink signal compression, giving particular attention to scenarios where the MIMO capacity is fundamentally limited by the available fronthaul capacity. The main contributions of Chapter 4 are:

- An upper bound on the achievable sum capacity when the sampled signals from each antenna are forwarded directly over fronthaul, demonstrating the need for more sophisticated data compression.
- Showing that with a single MIMO receiver, transform coding can asymptotically achieve the cut-set bound at high SNR – perfectly utilising the available fronthaul.
- A transform coding scheme for a single MIMO receiver that captures many of the fundamental benefits of massive MIMO, even when the overall MIMO capacity is physically limited by the fronthaul capacity.
- A scalable transform coding scheme for distributed MIMO that uses jointly optimised rate allocations to efficiently compress the received signals at different receivers.
- Adaptations of the proposed schemes to account for the use of imperfect CSI at the receivers.

Section 4.1.1 summarises these contributions in more detail.

Chapter 5 investigates the use of distributed dimension reduction for efficient signal compression in distributed MIMO C-RAN systems with multi-antenna radio heads and a large overall excess of BS antennas. Deploying a large excess of BS antennas in a distributed MIMO network brings well established benefits, but also a proportionate increase in signalling data. Dimension reduction techniques, which exploit sparsity in high dimensional signals to produce accurate lower dimension representations, are a common feature of many data compression schemes, but their explicit use for fronthaul data compression in distributed MIMO networks has not previously been studied in detail.

This chapter investigates the use of distributed dimension reduction for reducing the dimensionality of the fronthaul signal data on both the distributed MIMO uplink and downlink. The main contributions of Chapter 5 are:

- Showing that distributed dimension reduction of the uplink signals can be achieved by applying an optimised linear dimension reduction filter to the multi-antenna signal at each remote receiver.
- Showing that the dimension reduction filters that maximise the joint mutual information between the reduced dimension signals and the transmitted user symbols are a truncated form of the conditional Karhunen-Loeve transform, found using a block coordinate ascent procedure.
- Numerical results demonstrating that significant dimension reduction can be achieved whilst preserving the key characteristics of the high dimension signals.

- A second, low complexity dimension reduction scheme based on matched filtering that can achieve efficient dimension reduction whilst also reducing the fronthaul data overheads related to the transfer of CSI.
- Showing that at high SNR, the use of distributed dimension reduction followed by simple lossy scalar compression results in an uplink MIMO sum capacity that scales approximately linearly with the available fronthaul capacity, and inversely with the signal dimension.
- Numerical results demonstrating that this dimension reduction-based uplink signal compression can be a highly efficient fronthaul compression strategy.
- A 'dual' two-stage downlink precoding scheme in which an inner precoder at the CP produces a set of low dimension signals which are quantized and transferred over fronthaul to the remote transmitters, where they are beamformed using a larger number of antennas.
- A downlink max-min power allocation scheme that uses the user power allocations to mitigate the impact of the quantization noise received by the users.

Section 5.1.1 summarises these contributions in more detail.

Chapter 6 provides a high level summary of the key findings of this work, and suggests some directions for further investigation.

1.4 Publications

The following papers were published during the course of this PhD:

1. F. Wiffen, M. Z. Bocus, A. Doufexi and A. Nix, "Phase-Only OFDM Communication for Downlink Massive MIMO Systems," *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, Porto, 2018, pp. 1-5 [222].
2. F. Wiffen, L. Sayer, M. Z. Bocus, A. Doufexi and A. Nix, "Comparison of OTFS and OFDM in Ray Launched sub-6 GHz and mmWave Line-of-Sight Mobility Channels," *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Bologna, 2018, pp. 73-79 [224].
3. F. Wiffen, M. Z. Bocus, A. Doufexi and A. Nix, "Distributed MIMO Uplink Capacity Under Transform Coding Fronthaul Compression," *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, Shanghai, China, 2019, pp. 1-6 [223].
4. F. Wiffen, M. Z. Bocus, A. Doufexi and W. H. Chin, "MF-based Dimension Reduction Signal Compression for Fronthaul-Constrained Distributed MIMO C-RAN," *2020 IEEE Wireless Communications and Networking Conference (WCNC)*, Seoul, Korea (South), 2020, pp. 1-8 [221].

Papers 1, 3 & 4 correspond, respectively, to the topics discussed in Chapters 3, 4 & 5. However, many of the **key findings** outlined in this thesis are **currently unpublished**, due to recent progess and time constraints. It is hoped that the unpublished material from Chapters 3 & 4 could be published in 2 additional conference papers, and the material from Chapter 5 published as a journal paper. This is discussed in more detail in Sections 3.1.2, 4.1.2 & 5.1.2

Entry 2 was also published under the funding provided for this PhD, but addresses a separate wireless communications topic unrelated to MU-MIMO, and has been omitted from this thesis for reasons of space and coherence.

Chapter 2

Fundamentals of Multi-User MIMO Communication

This chapter outlines the background theory, concepts, techniques and models that underpin multi-user MIMO (MU-MIMO) technology – providing a basis for understanding the research material contained in subsequent chapters.

The discussion begins by considering single antenna communication systems. A mathematical model for the linear time invariant wideband wireless channel is first given, before single carrier and OFDM waveforms and the ideas of channel capacity and coding are introduced. The wireless fading channel is then defined, and concepts of ergodic capacity, outage capacity and diversity introduced.

These ideas are then extended to multi-antenna communication. The use of multiple antennas at one side of a wireless link in SIMO and MISO systems to improve link reliability is discussed. It is then shown that the use of multiple antennas at a base station enables multiple users to be spatially multiplexed on the same time-frequency resource. Uplink and downlink multi-user MIMO channels are defined, and the benefits to system capacity provided by spatial multiplexing are shown.

The next section focuses on the MIMO signal processing techniques required to spatially multiplex users, with particular attention given to linear detection and precoding methods, which represent a practical and scalable solution for cellular systems.

Two simple channel models – ray-based and correlated Rayleigh fading – that capture spatial characteristics of the MU-MIMO channel are then outlined. Channel estimation techniques for obtaining the channel state information required for precoding and detection are discussed and a model for assessing the impact of estimation errors on capacity is established. The use of power control algorithms in cellular MU-MIMO systems is motivated, and appropriate schemes outlined.

Finally, two advanced MU-MIMO architectures that are expected to play a significant role in future cellular & wireless networks are introduced – massive MIMO & distributed MIMO. The key features of these architectures are described and their benefits discussed, before some outstanding areas of research are outlined – paving the way for the research provided in Chapters 3 to 5.

Whilst aiming to provide an introduction to important modern digital communication techniques, the scope and length of discussion in this chapter is restricted to focus mainly on those ideas used by later research material, and therefore some concepts relevant to the design and analysis of wireless MIMO communications systems – for example space-time coding and diversity-multiplexing trade-offs – are omitted, or given only a cursory survey.

2.1 Wireless Digital Communication

The fundamental aim of a digital communication system is the error-free transfer of a set of binary information bits, b_i, over a physical channel, subject to an appropriate set of quality of service constraints, such as transmission rate or delay. The groundbreaking work of Claude Shannon in his seminal *Mathematical Theory of Communication* provides a mathematical framework for analysing the performance and limits of such systems, whilst providing crucial insights into the architectures required to approach these limits [191].

Shannon showed that reliable communication can be achieved using a process called *channel coding*, where strings of information bits are mapped to appropriately chosen strings (or codewords) of symbols, s_n, which are transmitted into the channel. At the receiver, a reverse decoding process is then used to estimate the information bits from the noisy or distorted symbols at the channel output. Providing the coding system is properly designed and information is transmitted at a rate not greater than the *channel capacity*, the original bits can be recovered with an arbitrarily small probability of error.

In modern wireless digital communication systems, coded symbols are transmitted through a wireless channel by first mapping them to a continuous time baseband signal, $x(t)$, through baseband modulation and digital-to-analogue conversion. The baseband signal is then upconverted to an appropriate radio frequency, using an IQ modulator, after which it is amplified and radiated by an antenna into the wireless channel. At the receiver, a second antenna is excited by this electro-magnetic radiation, and the reverse process takes place, as shown in Figure 2.1.

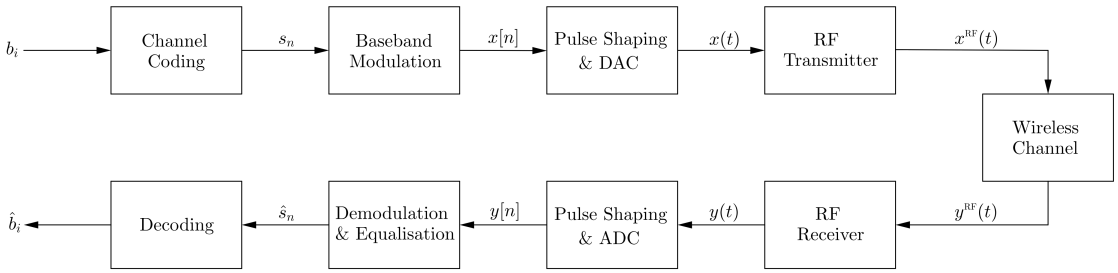

Figure 2.1: Basic wireless communication system block diagram.

2.1.1 The Wireless Channel

The research in this thesis focuses on techniques that operate in the digital domain, and hence it is necessary to develop an equivalent channel model that describes the relationship between the sampled digital signals at transmitter and receiver.

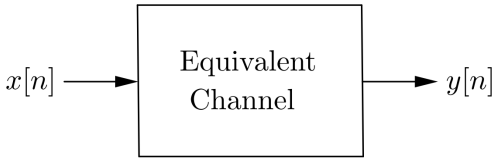

Figure 2.2: Equivalent digital channel.

A static wireless channel with real valued RF transmit and receive signals $x_{\mathrm{RF}}(t) \in \mathbb{R}$ and $y_{\mathrm{RF}}(t) \in \mathbb{R}$ is a noisy linear time invariant (LTI) system with passband impulse response $h_{\mathrm{RF}}(t) \in \mathbb{R}$,

$$\begin{aligned} y_{\mathrm{RF}}(t) &= h_{\mathrm{RF}}(t) * x_{\mathrm{RF}}(t) + \eta_{\mathrm{RF}}(t) \\ &= \int h_{\mathrm{RF}}(\tau) x_{\mathrm{RF}}(t-\tau) d\tau + \eta_{\mathrm{RF}}(t), \end{aligned} \tag{2.1}$$

where $\eta_{\mathrm{RF}}(t)$ is complex additive white Gaussian noise (AWGN) [174]. It can be shown that the baseband input-output relation is also a (complex) noisy LTI system

$$\begin{aligned} y(t) &= h(t) * x(t) + \eta(t) \\ &= \int h(\tau) x(t-\tau) d\tau + \eta(t), \end{aligned} \tag{2.2}$$

with complex input signal $x(t) \in \mathbb{C}$, received signal $y(t) \in \mathbb{C}$, AWGN $\eta(t) \in \mathbb{C}$ and baseband equivalent channel impulse response $h(t) \in \mathbb{C}$ – which captures the combined effect of the passband wireless channel, up-/down-conversion, and all linear components, such as filters, on the transmit and receive signal paths.

The continuous transmit signal is generated from a sequence of digital samples, $x[n] \in \mathbb{C}$ with sample spacing T_s seconds, filtered with a transmit pulse[1], $g_{tx}(t)$,

$$x(t) = \sum_n \int x[n] \delta(t - nT_s - \tau) g_{tx}(\tau) d\tau \tag{2.3}$$

$$= \sum_n x[n] g_{tx}\big(t - nT_s\big). \tag{2.4}$$

Similarly, at the receiver the continuous receive signal is filtered using a receive pulse, $g_{rx}(t)$, and sampled at time intervals nT_s, to give digital receive signal

$$y[n] = \int g_{rx}(t' - t) y(t') dt' \Big|_{t=nT_s}. \tag{2.5}$$

The overall relationship between digital input and output signals can be described by an equiv-

[1]Pulse shaping is used to limit the spectrum of the signal, and is in practice often carried out digitally through oversampling – this does not affect the overall equivalent channel expression.

alent digital impulse response, $h[m] \in \mathbb{C}$,

$$\begin{aligned} y[n] &= \sum_m h[m]x[n-m] + \eta[n] \\ &= h[n] * x[n] + \eta[n], \end{aligned} \tag{2.6}$$

and equivalent noise $\eta[n]$. This impulse response captures the combined effect of the baseband equivalent channel and transmit and receive filters

$$h[n] = h(t) * g_{rx}(t) * g_{tx}(-t) \Big|_{t=nT_s}, \tag{2.7}$$

and the discrete time LTI transfer function fully describes the wireless channel as seen by the digital communication system. The specific structure of the transfer function depends on the propagation environment through which the signal passes. In practical wireless channels, all propagation paths between transmitter and receiver have a finite physical length, and the channel transfer function therefore has finite support, L,

$$h[n] = \begin{cases} h[n], & n \in [0, L-1] \\ 0, & \text{else.} \end{cases} \tag{2.8}$$

Channels with multiple non-zero entries (taps), $L > 1$, are called *wideband*. Most practical high bandwidth communication systems operate in wideband channels, since they operate at high sampling rates (small T_s) in multi-path propagation environments. When the channel has a single tap ($L = 1$), it is called *narrowband*, and the transfer function reduces to a simple multiplication

$$y[n] = h[0]x[n] + \eta[n]. \tag{2.9}$$

2.1.2 Modulation

Digital communication requires a mapping from the transmission symbols, s_n, to the output samples, $x[n]$. A large number of digital modulation schemes exist that have been implemented across a range of wireless technologies. European second generation cellular systems employ a form of phase shift keying, which produced a signal with constant envelope that allowed for very power efficient amplification [151]. American second generation and all third generation cellular systems use direct sequence spread spectrum modulation methods, in which the symbols are spread across a large signal bandwidth in order to capture additional transmission *diversity* and increase reliability [67]. Fourth and fifth generation systems, as well as current WiFi standards, have all adopted multi-carrier modulation schemes based on orthogonal frequency division multiplexing (OFDM), due to its flexibility and high spectral efficiency [41].

This section first outlines simple single carrier modulation, before the operation of the OFDM waveform, as well as simple single carrier modulation.

Single Carrier

In simple single carrier modulation the symbols are directly mapped to output samples [55],

$$x[n] = s_n, \tag{2.10}$$

resulting in a received signal

$$y[n] = \sum_{m=0}^{L-1} h[m]s_{n-m} + \eta[n]. \tag{2.11}$$

In wideband channels each receive sample contains the transmit symbol corresponding to the current sample index, plus a linear combination of the $L-1$ previous transmit symbols. If the taps $h[1], \ldots, h[L-1]$ are significant an *equaliser* must be used to remove this inter-symbol interference (ISI). The topic of equaliser design has received extensive research attention and a large variety of architectures have been proposed – see e.g. [173], [213] & [55].

OFDM

Orthogonal frequency division multiplexing (OFDM) uses a cyclic prefix and pre- and post-processing stages based on the discrete/fast Fourier transform (FFT) to convert the wideband channel into a set of N parallel narrowband channels, over which the information symbols are transmitted [246]. This eliminates ISI, and makes channel equalisation trivial.

At the transmitter, a block of N symbols are precoded using the inverse FFT

$$s[n] = \frac{1}{\sqrt{N}} \sum_{l=0}^{N-1} s_l e^{j2\pi \frac{ln}{N}}. \tag{2.12}$$

This multiplexes symbols by using each one to modulate an orthogonal complex sinusoidal basis function, $e^{j2\pi \frac{ln}{N}}$. The composite signal contains N modulated *subcarriers* – OFDM is a multicarrier modulation scheme.

An OFDM symbol of length $N + N_{\text{cp}}$ samples is formed by appending a cyclic prefix to the N multiplexed symbols,

$$x[n] = \begin{cases} s[n+N], & n \in [-N_{\text{cp}}, -1] \\ s[n] & n \in [0, N-1]. \end{cases} \tag{2.13}$$

Providing the cyclic prefix is longer than the channel impulse response, $N_{\text{cp}} \geq L$, this converts the channel convolution into a cyclic convolution

$$y[n] = h[n] * x[n] + \eta[n] \tag{2.14}$$

$$= h[n] \circledast s[n] + \eta[n]. \tag{2.15}$$

At the receiver, the cyclic prefix is discarded and an FFT applied

$$y_l = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} y[n] e^{-j2\pi \frac{ln}{N}}. \tag{2.16}$$

Since the FFT converts a discrete circular convolution into a set of multiplications, the received signal on each subcarrier is

$$y_l = h_l s_l + \eta, \tag{2.17}$$

and each symbol is effectively passed through a separate narrowband channel with gain h_l, given by

$$h_l = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} h[n] e^{-j2\pi \frac{ln}{N}}. \tag{2.18}$$

A channel impulse response with $L = 12$ taps, and the associated subcarrier gains for OFDM symbol length $N = 128$ are shown in Figure 2.3. The presence of multiple taps in the impulse response causes deep 'fades' to occur on some subcarriers, e.g. subcarrier 10 in Figure 2.3. To prevent high symbol error rates, efficient coding of information across blocks of subcarriers is necessary [185], as discussed in Section 2.1.4.

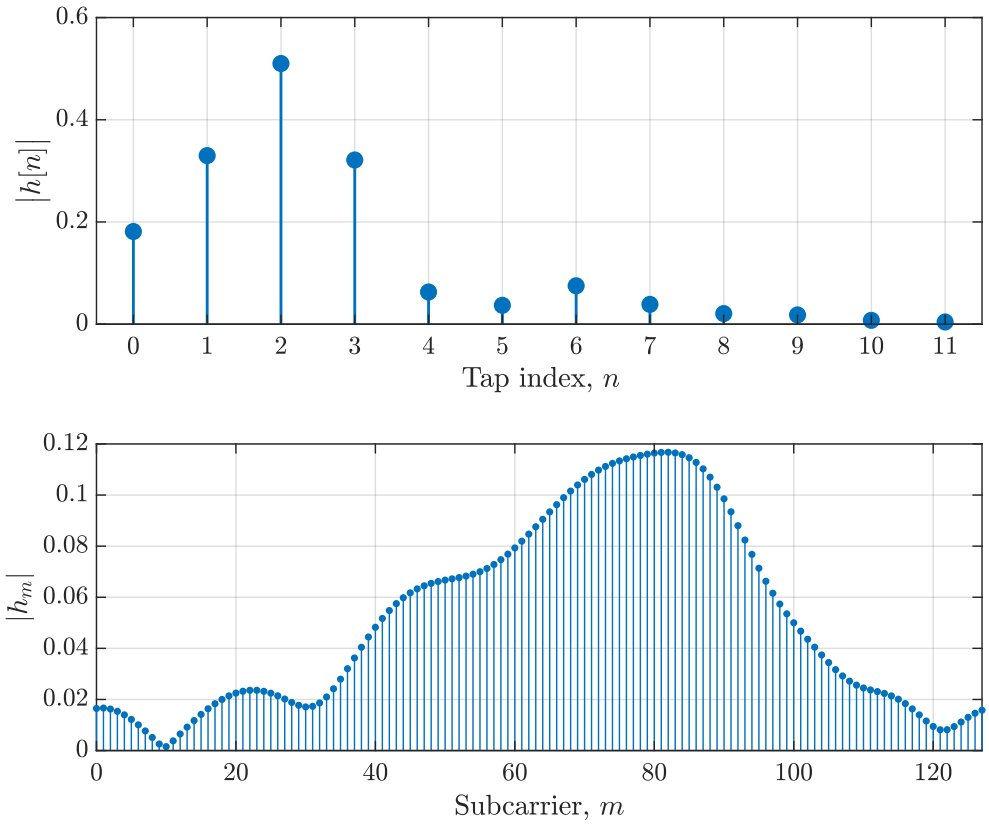

Figure 2.3: Channel impulse response (top) and associated OFDM subcarrier gains with $L = 12$, $N = 128$.

Since the FFT can be implemented very efficiently in hardware [241], OFDM is an attractive way of reducing the complexity of symbol equalisation. The conversion of the wideband channel into narrowband channels also simplifies analysis and makes OFDM a natural companion to MIMO technology [201]. Orthogonal frequency division multiple access (OFDMA) – in which different blocks of subcarriers are allocated to different users – is an efficient means of orthogonally multiplexing users [36], and a key component of fourth and upcoming fifth generation

cellular systems. There are a number of other benefits to OFDM [97] – such as its spectral properties and capacity achieving performance – that go beyond the scope of this discussion.

A drawback of OFDM is the overhead associated with transmitting the cyclic prefix, which reduces the overall spectral efficiency by a factor $N/(N+N_{\text{cp}})$. This can be an issue in channels with long impulse response (larger N_{cp} required), or when short OFDM symbols are used (small N). Another important issue is the high peak-to-average power ratio (PAPR) of the OFDM signal, which necessitates the use of expensive and inefficient linear power amplifiers [190]. Reduction of the PAPR of OFDM signals in MIMO systems is the focus of the research in Chapter 3.

2.1.3 Channel Capacity

One of the landmark results from Shannon's work on information theory was a means of calculating the maximum rate at which information can be transferred, error-free, through a communication channel [191]. He called this the capacity of the channel, $\mathcal{C}$, and showed that it is given by the mutual information between the receive signal and transmit symbols, $\mathcal{I}(y;s)$, maximised over the transmit symbol distribution, $p_{\text{s}}(s)$,

$$\mathcal{C} = \max_{p_{\text{s}}(s)} \mathcal{I}(y;s), \tag{2.19}$$

where

$$\mathcal{I}(y;s) = \mathcal{H}(y) - \mathcal{H}(y|s) \tag{2.20}$$

$$= \mathcal{H}(s) - \mathcal{H}(s|y) \tag{2.21}$$

with $\mathcal{H}(a)$ the entropy of the random variable, a, and $\mathcal{H}(a|b)$ the conditional entropy of a given b [47].

For the narrowband AWGN single input single output (SISO) channel (also referred to as the Gaussian channel),

$$y = hs + \eta, \tag{2.22}$$

with $\eta \sim \mathcal{CN}(0,1)$, this is given by

$$\mathcal{C} = \max_{p_{\text{s}}(s)} \mathcal{H}(hs+\eta) - \mathcal{H}(\eta). \tag{2.23}$$

Under a transmit power constraint, $\mathbb{E}[ss^{\dagger}] \leq \rho$, the capacity is maximised by setting $p_{\text{s}}(s)$ to a circularly symmetric Gaussian distribution, $s \sim \mathcal{CN}(0,\rho)$, giving

$$\mathcal{C} = \log_2\left(1+\rho|h|^2\right) \tag{2.24}$$

bits per channel use (bpcu).

Mapping strings of information bits, b_i, to strings/codewords of transmit symbols drawn randomly from a Gaussian distribution, $s_n \sim \mathcal{CN}\big(0,\rho\big)$, Shannon showed that this result holds asymptotically as the codeword length approaches infinity, with $\mathcal{C}$ the average number of bits

of information conveyed by each symbol.

Shannon's random coding scheme is not practical to implement, but represents a useful upper bound on the performance of all other coding schemes. Recent research has shown that discrete signal constellations, such as quadrature amplitude modulation (QAM), when used in conjunction with high performance error correcting codes, such as turbo [198], polar [157] or low-density parity check (LDPC) [177] codes, can come close to achieving this bound under finite coding block lengths, with practical computational complexity. Figure 2.4 shows that QPSK signalling with an off-the-shelf 3/5 rate LDPC code, codeword length 64800 bits, can achieve within 1 dB of the Shannon limit for the AWGN channel[2].

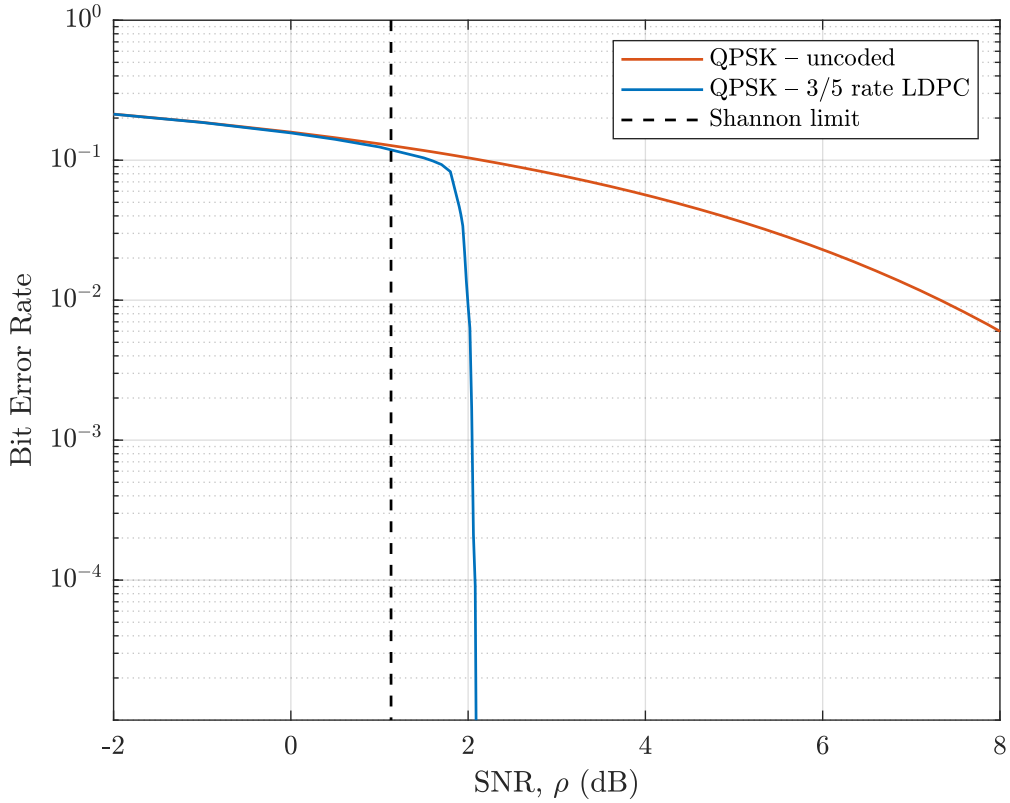

Figure 2.4: Bit error rate performance of QPSK with 3/5 rate LDPC coding.

The analysis of the capacity of systems under Gaussian signalling is therefore widely practised, since it is a mathematically tractable means of analysing the performance of real systems.

All Gaussian scalar channels have a capacity of the form

$$\mathcal{C} = \log_2(1 + \mathrm{SNR}) \tag{2.25}$$

where SNR is the *receiver* signal-to-noise ratio[3] (SNR $= \rho|h|^2$ for the AWGN channel). At high

[2]Optimised LDPC codes that come closer to the channel capacity have been demonstrated, e.g [179], [111].

[3]In this thesis, the term SNR is used to refer to either *transmit* SNR, $\frac{\mathbb{E}[|s|^2]}{\mathbb{E}[|\eta|^2]} = \rho$, or *receive* SNR, $\frac{\mathbb{E}[|hs|^2]}{\mathbb{E}[|\eta|^2]} = \rho|h|^2$. In many cases the two have a direct relationship; the distinction is made clear where important.

SNR, when $\rho|h|^2 \gg 1$, the capacity can be approximated

$$\mathcal{C} \approx \log_2 \rho + \log_2 \left(|h|^2\right), \tag{2.26}$$

and it is seen that capacity increases logarithmically with transmit power and channel gain, as shown in Figure 2.5.

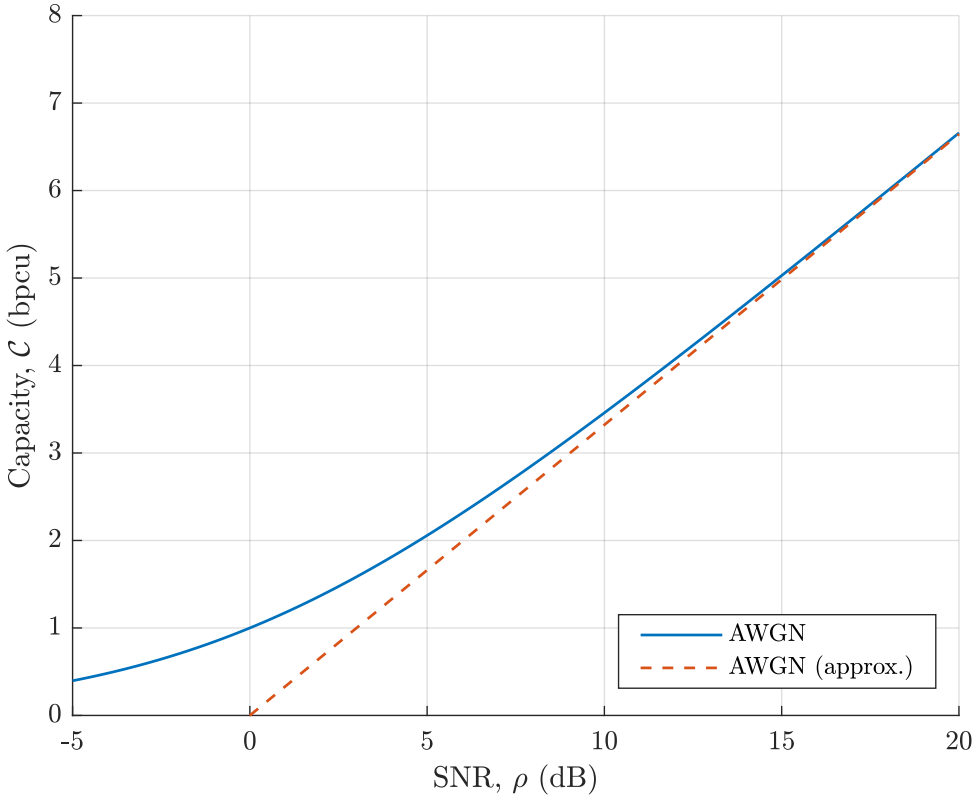

Figure 2.5: AWGN channel capacity, $|h|^2 = 1$.

Scalar Channel with Interference

The scalar channel with AWGN and additive interference, ϑ, is also frequently encountered in communication analysis

$$y = hs + \eta + \vartheta. \tag{2.27}$$

When the interference is Gaussian, $\vartheta \sim \mathcal{CN}(0, \sigma_\vartheta^2)$, the capacity is given by

$$\mathcal{C} = \log_2 \left(1 + \text{SINR}\right) \tag{2.28}$$

where SINR is the signal-to-noise-plus-interference ratio

$$\text{SINR} = \frac{\rho|h|^2}{\sigma_\vartheta^2 + 1} \tag{2.29}$$

$$= \frac{\text{signal power}}{\text{interference power} + \text{noise power}}. \tag{2.30}$$

Since Gaussian distributed interference represents a worst case interference [196], this is a lower bound on capacity for any interference with power σ_ϑ^2.

Throughput & Spectral Efficiency

The overall throughput of the channel is equal to the symbol rate, f_s, multiplied by the average number of bits per symbol, and can be upper bounded in terms of the channel capacity.

$$\text{throughput} = \text{symbol rate} \times \text{average bits per symbol} \tag{2.31}$$

$$\leq f_s \times \mathcal{C}. \tag{2.32}$$

The Shannon-Nyquist sampling theorem [192] states that the maximum symbol rate for a pass-band channel with bandwidth B is

$$f_s \leq B. \tag{2.33}$$

The spectral efficiency is the average number of bits transmitted per second per Hertz of bandwidth,

$$\text{SE} = \frac{\text{throughput}}{\text{bandwidth}, B} \quad \text{bps/Hz}, \tag{2.34}$$

and a measurement of how efficiently a communication systems uses the available bandwidth. As bandwidth becomes increasingly scarce, and therefore increasingly expensive for operators to use, spectral efficiency is a key performance metric for all modern wireless communication systems.

Substituting (2.31) and (2.33), the maximum spectral efficiency that can be achieved for a given channel and system is

$$\text{SE} \leq \mathcal{C} \quad \text{bps/Hz} \tag{2.35}$$

Since modern communication systems may achieve performance close to this spectral efficiency, the terms spectral efficiency and capacity are often used interchangeably. In practice, communication overheads, such as guard bands and control signals, mean that the actual information rate is lower than this.

2.1.4 Fading Channels

For most propagation environments, the wireless channel consists of multiple reflected/diffracted paths between transmitter and receiver, as illustrated in Figure 2.6.

In mobile environments, in which any of the receiver, transmitter or reflective objects within the propagation environment are in motion, the number of paths, their strengths and their delays change with time, and hence the wireless channel is a linear time variant system, with

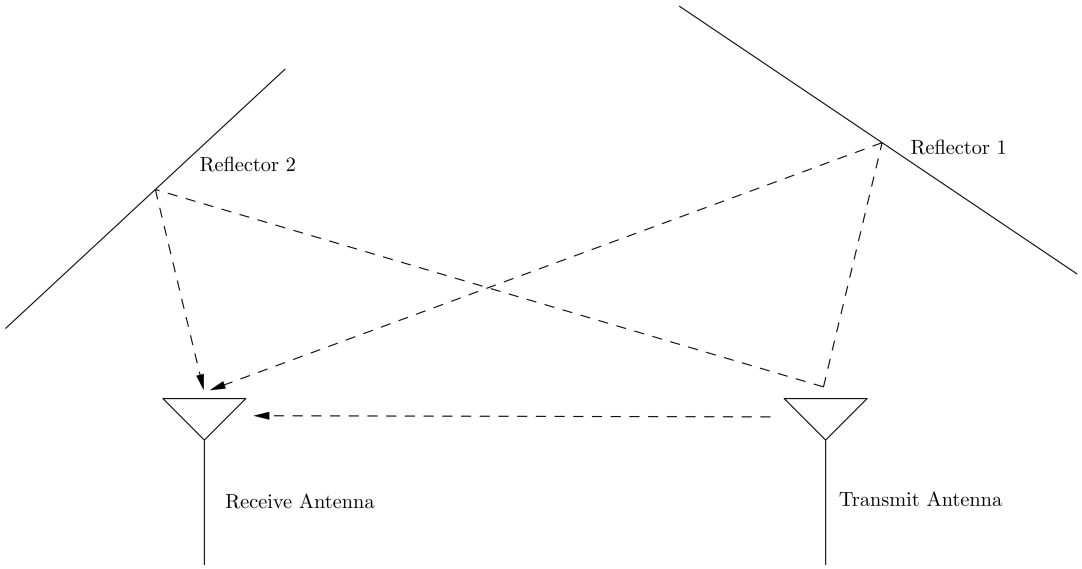

Figure 2.6: Wireless channel with direct and two reflected paths.

an impulse response that varies continuously with time, $h(t, \tau)$. However, for tractability, it is often sufficient to treat the channel as fixed (invariant) over short time intervals – the block fading model [212]. The *coherence time* is the length of time for which the channel impulse response remains approximately fixed, or highly correlated.

Within the coherence period, the channel impulse response can be modelled (ignoring hardware effects) as

$$h(t) = \sum_i \alpha_i \delta(t - \tau_i) e^{-j2\pi f_c \tau_i} \tag{2.36}$$

where f_c is the carrier frequency and α_i and τ_i the gain and delay associated with propagation path i. For OFDM modulation with subcarrier frequency spacing Δf and ideal pulse shaping, the channel gain on subcarrier n is

$$h_n = \sum_i \alpha_i e^{-j2\pi f_c \tau_i} e^{-j2\pi n \Delta f \tau_i}, \tag{2.37}$$

varying with frequency, $n\Delta f$, as well as time [224]. When the spread of path delays is small then the channel gain will tend to vary gradually with frequency, whereas when it is large it will tend to vary more quickly [212]. Since the channel gains on adjacent subcarriers will tend to be similar, they are often treated as being constant across a frequency band of certain size – the *coherence bandwidth*.

If the coherence bandwidth of a channel covers N_f subcarriers, whilst the coherence time allows for the transmission of N_s OFDM symbols, $N_f \times N_s$ data symbols can be transmitted within a *coherence block*, each experiencing (approximately) the same channel gain.

Rayleigh Fading

Since f_c is typically large ($\sim$ GHz), small displacements of the transmitter or receiver ($\sim$ cm) are sufficient to cause large phase changes for each transmission path, whilst their gains remain effectively constant. Therefore, (2.37) may be modelled as a random variable

$$h = \sum_i \alpha_i e^{j\theta_i} \tag{2.38}$$

where the phases of the paths are uniformly distributed and independent, $\theta_i \sim U(0, 2\pi)$. At some positions, the multipath components will interfere constructively, and the channel gain become large, whilst at others they will destructively interfere and the channel gain will be small. According to the central limit theorem [75], when a large number of paths with similar gain exist, the channel gain may be modelled as a circularly symmetric complex normal random variable

$$h \sim \mathcal{CN}(0, \beta) \tag{2.39}$$

with an envelope that follows the Rayleigh distribution

$$|h| \sim \text{Rayleigh}(\sqrt{\beta/2}) \tag{2.40}$$

where β is the average received signal strength

$$\beta = \mathbb{E}\big[|h|^2\big] = \sum_i |\alpha_i|^2. \tag{2.41}$$

This Rayleigh fading channel model is widely used since it captures the dynamics of fading well and is analytically tractable – often enabling insightful closed form expressions to be found [144]. In scenarios where the underpinning assumptions do not hold, for example where there is a dominant (such as a line-of-sight) propagation path, fading models based on the Rician or Nakagami distributions may provide a better fit.

Pathloss Model

For modelling purposes, a value for β can be obtained using a distance dependent pathloss model. In the literature a variety of pathloss models have been proposed based on empirical measurement campaigns. A commonly used model, and the one used here in later chapters, is the log-distance path loss model [214]

$$\beta = 147.5 + G_T + G_R - 20\log_{10}(f_c) - 10\gamma\log_{10}(d) + \Psi \quad \text{(dB)} \tag{2.42}$$

where

- G_T and G_R are the gains of transmit and receive antennas in decibels.
- d is the distance between transmitter and receiver, in metres.
- γ is the pathloss exponent, typically in the range $\gamma \in [2, 5]$.

- Ψ is a normally distributed term, $\Psi \in \mathcal{N}(0, \sigma_\Psi^2)$, representing shadow fading due to, for example, blockage.

The parameters γ and σ_Ψ^2 depend on the type of propagation environment, and can be chosen as appropriate from empirical studies reported in the literature.

Slow Fading & Outage

Slow fading occurs when the coherence block size is large compared to the block of coded transmit symbols – for example in a channel that changes slowly. The data is then transmitted through a channel with capacity

$$\mathcal{C} = \log_2 \left(1 + \rho |h|^2\right), \tag{2.43}$$

where h is a single random fading channel gain realisation. Since the channel is constantly changing, the transmitter may not have knowledge of the instantaneous channel capacity, only the statistics of h, in which case it cannot choose a data transmission rate that it can guarantee will be decoded error-free. The outage probability, $p_{\text{out}}(\mathcal{C}^\star)$, is the likelihood that the channel cannot support a target transmission rate $\mathcal{C}^\star$ [212],

$$p_{\text{out}}(\mathcal{C}^\star) = \mathbb{P}\text{r}\left[\mathcal{C} < \mathcal{C}^\star\right], \tag{2.44}$$

and is a measure of the reliability of a communication system. Similarly, the ϵ-outage capacity, $\mathcal{C}_\epsilon$, is the maximum transmission rate that has outage probability of less than ϵ,

$$\mathcal{C}_\epsilon = \max_{\mathcal{C}^\star} \left[\mathcal{C}^\star : \; p_{\text{out}}(\mathcal{C}^\star) \leq \epsilon\right]. \tag{2.45}$$

At high SNR, the outage probability of the Rayleigh channel scales approximately inversely with SNR,

$$p_{\text{out}}(\mathcal{C}^\star) \sim \frac{1}{\rho}. \tag{2.46}$$

Diversity

To reduce the outage probability, signal *diversity* can be employed. In a system with diversity, each data symbol is transmitted over multiple fading channels, with the received signals jointly used for symbol detection. If the channels fade independently, then the probability of all the channel gains being small, and the system going into outage, is reduced.

For a system with a diversity gain of d, the outage probability scales like

$$p_{\text{out}}(\mathcal{C}^\star) \sim \rho^{-d} \tag{2.47}$$

at high SNR [124], as shown in Figure 2.7.

This can be achieved using repeated transmission of each symbol in d different coherence blocks – for example different time slots or different frequencies (time/frequency diversity). However, this repetition coding strategy comes at the cost of reduced overall achievable throughput, since the symbol transmission rate must be decreased by a factor of $1/d$.

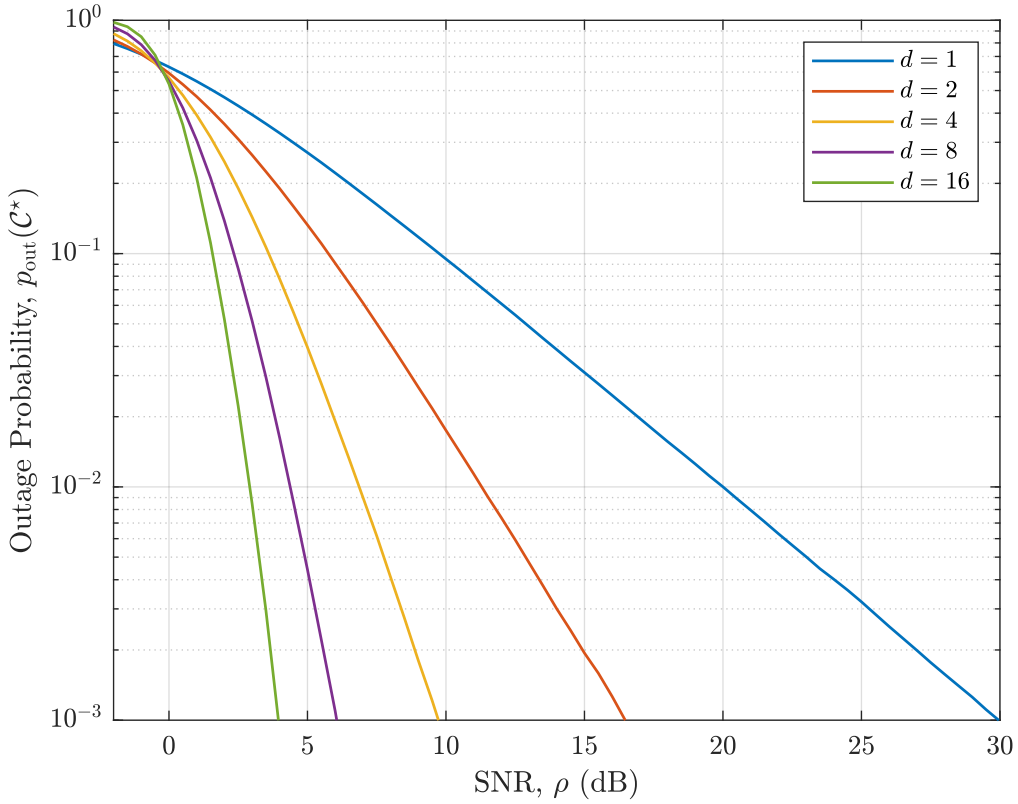

Figure 2.7: Outage probability for Rayleigh channel with different diversity levels, $\mathcal{C}^\star = 1$ bpcu.

Coding of data across subcarriers with OFDM [185] – *frequency diversity* – is a more efficient means of exploiting signal diversity, and is commonly used in many current wireless systems. The use of multiple antennas to exploit *spatial diversity* is another attractive technique, since it enables a diversity gain within a single time-frequency coherence block, by transmitting and/or receiving symbols with multiple antennas simultaneously.

Fast Fading

When the coherence block size is small compared to the length of data transmission, the channel can be said to experience *fast fading* [212], and each transmission of a block of coded symbols spans a number of coherence intervals. This can occur in systems with high mobility or wide-band systems employing OFDM. An important result from information theory says that as the number of coherence blocks spanned by transmission tends to infinity, an error free constant information rate of

$$\overline{\mathcal{C}} = \mathbb{E}\big[\mathcal{C}\big] \tag{2.48}$$

$$= \mathbb{E}\big[\log_2\left(1 + \rho|h|^2\right)\big] \tag{2.49}$$

can be achieved with appropriate symbol coding [70]. This is referred to as the *ergodic* capacity of the channel. According to Jensen's inequality [135], for the SISO channel the ergodic capacity

is upper bounded by the capacity of the SISO channel with average channel gain

$$\overline{\mathcal{C}} \leq \log_2\left(1 + \rho\mathbb{E}\left[|h|^2\right]\right). \tag{2.50}$$

Figure 2.8 shows the ergodic and 10%-outage capacity of the Rayleigh fading channel with $\beta = 1$.

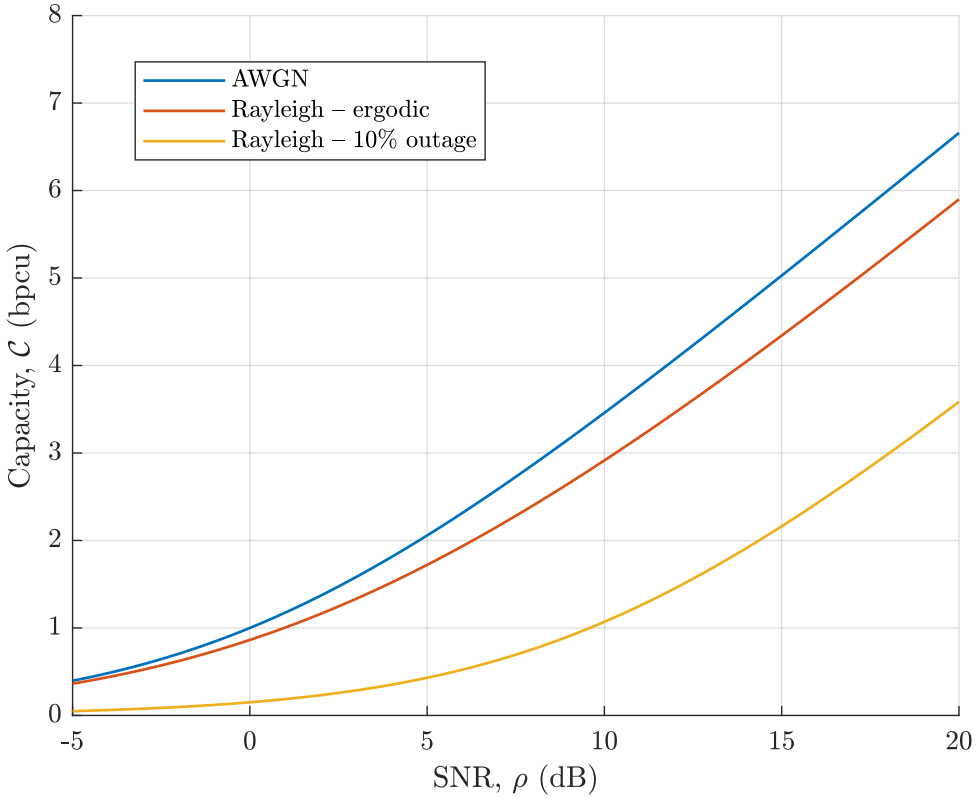

Figure 2.8: SISO channel capacity.

2.2 From SISO to MU-MIMO

The logarithmic scaling of capacity with transmit power limits the spectral efficiency that can be practically achieved in single input single output (SISO) wireless systems. The use of multiple antennas to enable multiple data streams to be multiplexed within a single channel use can overcome this limitation in capacity scaling, as well as providing improving link reliability by providing additional diversity.

This section introduces multi-user multiple input multiple output (MU-MIMO) technology, in which a base station equipped with multiple antennas serves multiple users on the same time-frequency resource – a technique also known as *space division multiple access* (SDMA).

The use of a multiple antenna BS to serve a single user is first considered, and shown to increase the link SNR whilst providing a diversity gain. This is then extended to a full MU-MIMO system and analysed in terms of uplink and downlink operation. Sum capacity

expressions are then derived which show that spatial multiplexing increases the *sum* spectral efficiency (SSE) by a factor of approximately K compared to SISO systems.

2.2.1 SIMO & MISO

Before studying the MU-MIMO channel it is instructive to first consider the SIMO (single input multiple output) and MISO (multiple input single output) channels, which are characterised by the use of multiple receive and transmit antennas, respectively.

SIMO

In a single input multiple output (SIMO) system, transmissions from a single transmit antenna are received simultaneously by M receive antennas, as shown in Figure 2.9.

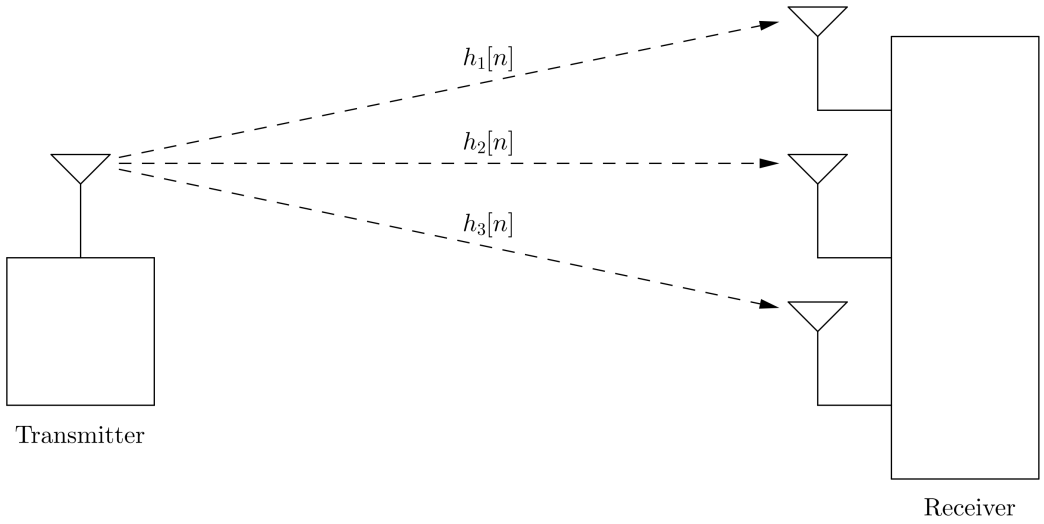

Figure 2.9: SIMO system with $M = 3$.

The received signal at receiver m is given by

$$y_m[n] = h_m[n] * x[n] + \eta_m[n], \tag{2.51}$$

where $x[n] \in \mathbb{C}$ is the transmit signal and $y_m[n]$ the signal received at antenna m after passing through a wideband propagation channel $h_m[n] \in \mathbb{C}$, as defined in Section 2.1.1.

The full received vector can be compactly written as a vector convolution

$$\begin{bmatrix} y_1[n] \\ \vdots \\ y_M[n] \end{bmatrix} = \begin{bmatrix} h_1[n] \\ \vdots \\ h_M[n] \end{bmatrix} * x[n] + \begin{bmatrix} \eta_1[n] \\ \vdots \\ \eta_M[n] \end{bmatrix} \tag{2.52}$$

or

$$\mathbf{y}[n] = \mathbf{h}[n] * x[n] + \boldsymbol{\eta}[n]. \tag{2.53}$$

If OFDM modulation is used, this set of M convolutions is transformed into a set of parallel vector multiplications

$$\mathbf{y}[n] = \mathbf{h}[n] * x_{\text{cp}}[n] + \boldsymbol{\eta}[n] \tag{2.54}$$

$$\implies \mathbf{y}_l = \mathbf{h}_l x_l + \boldsymbol{\eta}_l, \tag{2.55}$$

This transformation can significantly reduce signal processing complexity, and hence OFDM is a natural companion to SIMO, and other multiple antenna technologies [201].

The capacity of the SIMO narrowband channel,

$$\mathbf{y} = \mathbf{h}x + \boldsymbol{\eta} \tag{2.56}$$

is given, similarly to (2.19), by

$$\mathcal{C}_{\text{SIMO}} = \max_{p_{\text{x}}(x)} \mathcal{I}(\mathbf{y}; x) \tag{2.57}$$

$$= \max_{p_{\text{x}}(x)} \mathcal{H}(\mathbf{h}x + \boldsymbol{\eta}) - \mathcal{H}(\boldsymbol{\eta}). \tag{2.58}$$

Assuming independent noise with unit variance[4], $\boldsymbol{\eta} \sim \mathcal{CN}(0, \mathbf{I}_M)$, capacity is again maximised by Gaussian signalling [212], $x \sim \mathcal{CN}(0, \rho)$,

$$\mathcal{C}_{\text{SIMO}} = \log_2\left(1 + \rho\|\mathbf{h}\|^2\right) \tag{2.59}$$

$$= \log_2\Big(1 + \rho \sum_{m=1}^{M} |h_m|^2\Big). \tag{2.60}$$

This capacity is achieved by applying a matched filter (MF) [187] at the receiver

$$\tilde{y} = \mathbf{h}^\dagger \mathbf{y} \tag{2.61}$$

$$= \|\mathbf{h}\|^2 x + \mathbf{h}^\dagger \boldsymbol{\eta} \tag{2.62}$$

to produce a scalar channel, as in Section 2.19, with channel gain $\|\mathbf{h}\|^2$ and receive SNR $\rho\|\mathbf{h}\|^2$. The receive matched filter is also known as maximum ratio combining (MRC) [24]. Full knowledge of the channel realisation $\mathbf{h}$ is not required at the transmitter to achieve this capacity.

Under channel fading, the outage probability and ergodic capacity are defined as (2.44) & (2.48). Providing the receive antennas are sufficiently spaced, the channel impulse responses and hence signals received at each antenna will differ (see Section 2.4.1). When the elements of $\mathbf{h}$ are independently Rayleigh fading, a diversity gain of M is achieved. Furthermore, since

$$\mathbb{E}\big[\|\mathbf{h}\|^2\big] = M \times \mathbb{E}\big[|h_m|^2\big], \tag{2.63}$$

a received SNR gain, or *array gain*, of M is also achieved relative to SISO communication. This is illustrated in the capacity CDF in Figure 2.10, with the array gain producing an increase in the mean capacity whilst the diversity gain reduces capacity variations.

[4]This may be enforced by appropriate application of a whitening transform to $\mathbf{y}$.

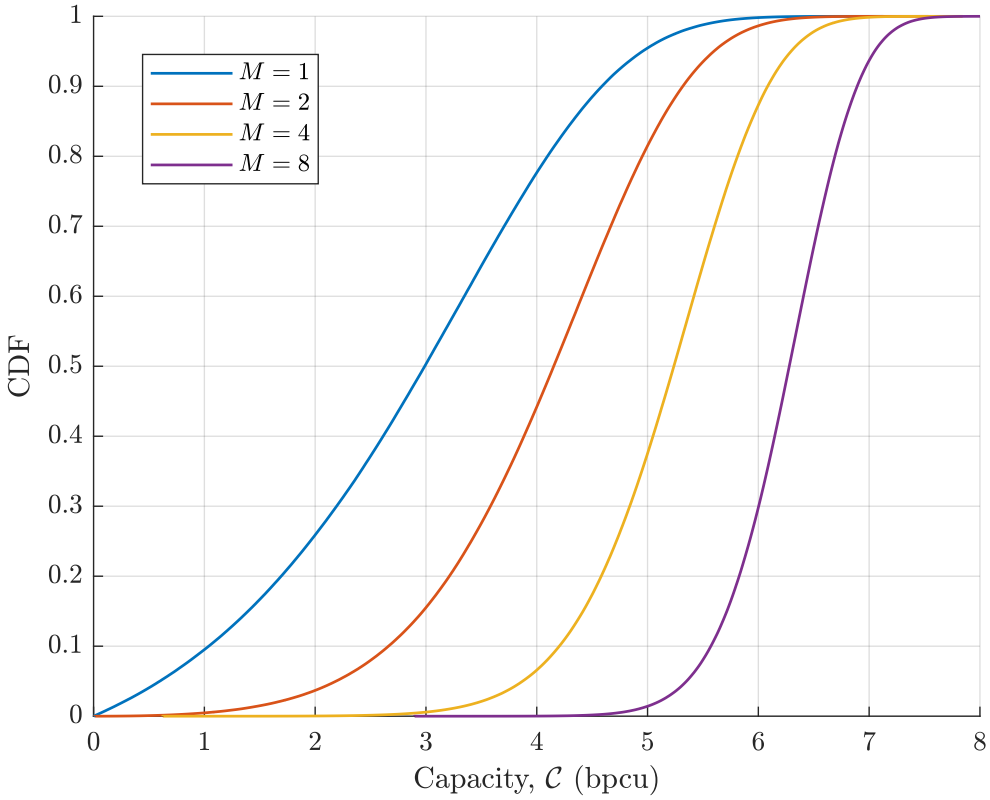

Figure 2.10: CDF of SIMO channel capacity, $\rho = 10$ dB

MISO

Reversing the roles of transmit and receive antennas, so that multiple antennas now – co-operatively – transmit to a single receive antenna, as shown in Figure 2.11, results in the multiple input single output channel (MISO), the dual of the SIMO channel.

The narrowband MISO channel is given by

$$y = \mathbf{h}^T \mathbf{x} + \eta, \tag{2.64}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1; & \dots & x_M \end{bmatrix} \tag{2.65}$$

are the transmit symbols at each antenna and

$$\mathbf{h} = \begin{bmatrix} h_1; & \dots & h_M \end{bmatrix} \tag{2.66}$$

the respective narrowband channel gains.

Assuming the channel realisation, $\mathbf{h}$, is known to the transmitter, the capacity-maximising transmit signal is found by precoding a data symbol, $s \sim \mathcal{CN}(0, \rho)$, using the matched filter,

$$\mathbf{x} = \frac{\mathbf{h}^*}{\|\mathbf{h}\|} s, \tag{2.67}$$

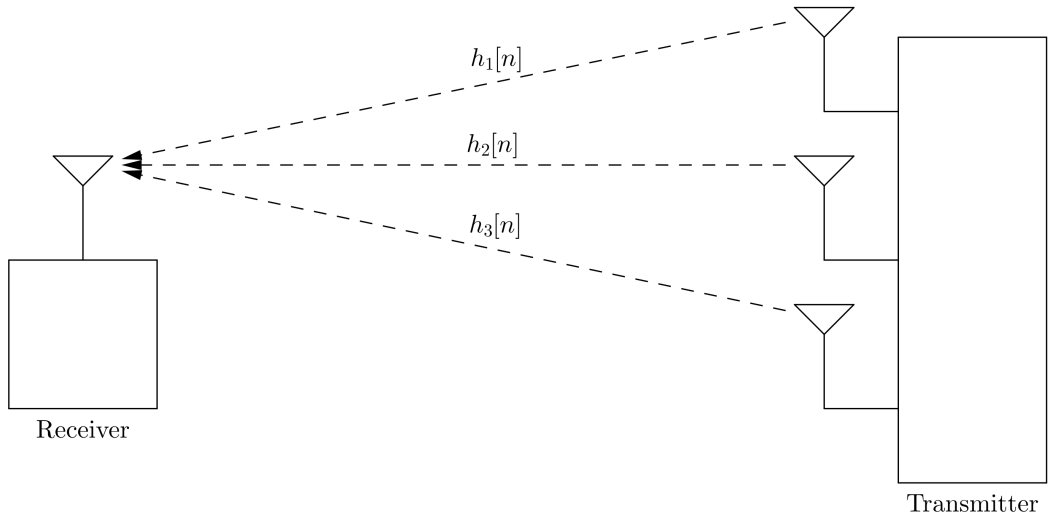

Figure 2.11: MISO system with $M = 3$.

where the denominator ensures a total power constraint, ρ, is met. This is also known as maximum ratio transmission (MRT). As with the receive matched filter, MRT results in a scalar channel,

$$y = \frac{\|\mathbf{h}\|^2}{\|\mathbf{h}\|} s + \eta, \tag{2.68}$$

with $\text{SNR} = \rho\|\mathbf{h}\|^2$ and capacity

$$\mathcal{C}_{\text{MISO}} = \log_2\left(1 + \rho\|\mathbf{h}\|^2\right) \tag{2.69}$$

$$= \log_2\left(1 + \rho \sum_{m=1}^{M} |h_m|^2\right). \tag{2.70}$$

Under the same total transmit power constraint, the capacity of the SIMO and MISO channels with transmitter channel knowledge are identical. This is in fact a result from the more general theory of uplink-downlink duality – a useful tool for relating the capacities of uplink and downlink dual channels [218].

In reality, obtaining the CSI at the transmitter required for data precoding is non-trivial, as discussed in Section 2.4.2. When CSI is not available a space time code can instead be used to achieve a diversity gain. For example, when $M = 2$, the well known Alamouti scheme also achieves a diversity gain of $d = 2$ without requiring channel knowledge at the transmitter, but does not achieve an array gain [3].

2.2.2 MU-MIMO

SIMO and MISO technology can be implemented on the uplink and downlink of a cellular system, respectively, by employing multiple antennas at the base station and a single antenna at the user. Whilst improving link reliability through diversity, the logarithmic scaling of capacity with transmit power means the sum capacity improvements achieved by SIMO/MISO

systems are modest. The simultaneous transmission, or *spatial multiplexing*, of multiple user data streams offers a means of overcoming these capacity limitations in multi-antenna systems.

Uplink

On the multi-user multiple input multiple output (MU-MIMO) uplink with K single antenna users simultaneously transmitting, the received signal at the BS is the superposition of K SIMO channels,

$$\mathbf{y}^{\mathrm{UL}} = \sum_{k=1}^{K} \sqrt{p_k}\mathbf{h}_k^{\mathrm{UL}} x_k^{\mathrm{UL}} + \boldsymbol{\eta} \tag{2.71}$$

where $\mathbf{h}_k^{\mathrm{UL}}$ and x_k^{UL} are the channel vector and transmit symbol for user k, respectively, with user power control coefficients, p_k, introduced. This is illustrated in Figure 2.12 for two users.

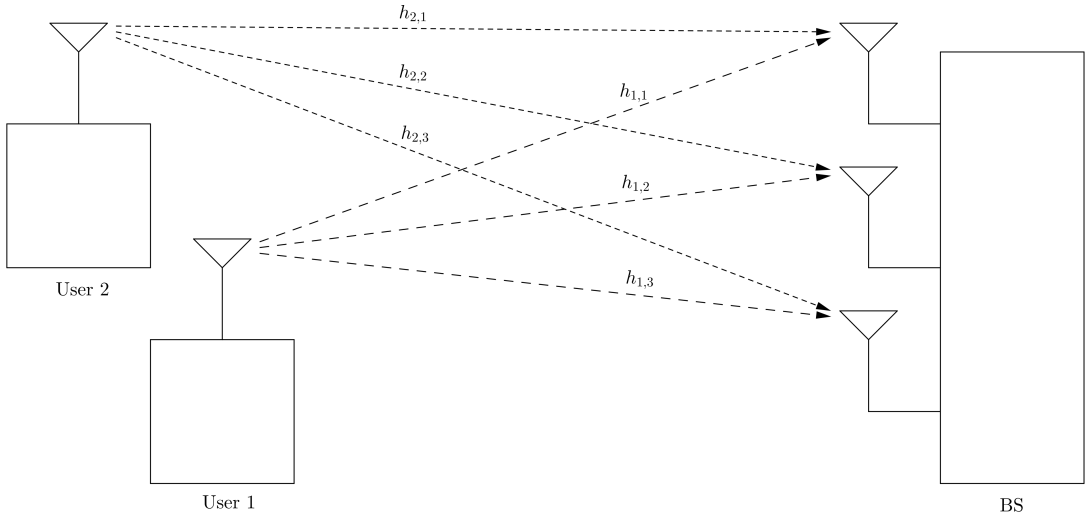

Figure 2.12: MU-MIMO system with $M = 3$ and $K = 2$.

It is often convenient to write (2.71) in matrix form

$$\mathbf{y}^{\mathrm{UL}} = \mathbf{H}^{\mathrm{UL}}\mathbf{P}^{1/2}\mathbf{x}^{\mathrm{UL}} + \boldsymbol{\eta}, \tag{2.72}$$

where

$$\mathbf{H}^{\mathrm{UL}} = \begin{bmatrix} \mathbf{h}_1^{\mathrm{UL}} & \dots & \mathbf{h}_K^{\mathrm{UL}} \end{bmatrix}, \tag{2.73}$$

$$\mathbf{x}^{\mathrm{UL}} = \begin{bmatrix} x_1^{\mathrm{UL}} & \dots & x_K^{\mathrm{UL}} \end{bmatrix}, \tag{2.74}$$

and

$$\mathbf{P} = \mathrm{diag}\big(p_1, \dots, p_K\big). \tag{2.75}$$

The received signal, $\mathbf{y}^{\mathrm{UL}}$, contains received transmissions from all K users, and the BS must apply *multi-user detection* to estimate each user transmit symbol and decode its message. In a cellular environment, the spatial separation of different users means that the channel vectors can generally be assumed to vary independently, and hence it is reasonable to assume that any set of $K \leq M$ user vectors are linearly independent [212]. The received signals (2.71) & (2.72)

therefore constitute a linear set of equations with M observations in K unknowns, and so for $K \leq M$ it is possible for the BS to estimate each x_k^{UL}.

Downlink

Similarly, the MU-MIMO downlink can be thought of as a set of K MISO channels, where user k receives the BS transmit vector, $\mathbf{x}^{\mathrm{DL}}$, passed through its downlink channel, $\mathbf{h}_k^{\mathrm{DL}}$,

$$y_k^{\mathrm{DL}} = (\mathbf{h}_k^{\mathrm{DL}})^T \mathbf{x}^{\mathrm{DL}} + \eta \tag{2.76}$$

In this case, the transmit vector must be chosen, or *precoded*, at the BS such that each user receives its desired stream of data, with minimal interference from the other streams. Writing the (2.76) in matrix form,

$$\mathbf{y}^{\mathrm{DL}} = \mathbf{H}^{\mathrm{DL}} \mathbf{x}^{\mathrm{DL}} + \boldsymbol{\eta}, \tag{2.77}$$

gives a matrix with K known values, and M variables. Again, providing $M \geq K$, an $\mathbf{x}^{\mathrm{DL}}$ can be found to give the desired signal at the users. This precoding requires the BS to have knowledge of the downlink channel matrix $\mathbf{H}^{\mathrm{DL}}$.

When the uplink and downlink channels between each antenna pair are the same, the channel is called reciprocal, and

$$\mathbf{h}_k^{\mathrm{DL}} = \mathbf{h}_k^{\mathrm{UL}} \tag{2.78}$$

$$\mathbf{H}^{\mathrm{DL}} = \left(\mathbf{H}^{\mathrm{UL}}\right)^T. \tag{2.79}$$

This occurs when uplink and downlink transmissions take place within the same time-frequency coherence block, such as in time division duplex (TDD) systems, and has the benefit that channel state information can be obtained on the uplink and re-used for downlink precoding [134].

For ease of notation, from here on in the UL & DL channel superscripts are omitted, with transmission direction becoming clear from context.

2.2.3 Channel Capacity

The key benefit of MU-MIMO is its ability to achieve high system throughputs and sum spectral efficiencies through the spatial multiplexing of multiple users on the same time frequency resource. Capacity analysis of MIMO systems gives a valuable insight into the throughput gains that can be achieved by modern communication systems operating close to Shannon limits.

MU-MIMO Uplink

The sum capacity of the MU-MIMO uplink is

$$\mathcal{C}_{\mathrm{SUM}}^{\mathrm{UL}} = \max_{p_{\mathbf{x}}(\mathbf{x})} \mathcal{I}(\mathbf{y}; \mathbf{x}) \tag{2.80}$$

$$= \max_{p_{\mathbf{x}}(\mathbf{x})} \mathcal{H}(\mathbf{y}) - \mathcal{H}(\mathbf{y}|\mathbf{x}). \tag{2.81}$$

Assuming the symbols transmitted by each user are independent with $\mathbb{E}[\mathbf{x}\mathbf{x}^\dagger] = \rho\mathbf{I}_K$, the optimal symbol distribution is Gaussian [212], $x_k \sim \mathcal{CN}(0, \rho)$, giving

$$\mathcal{C}_{\text{SUM}}^{\text{UL}} = \log_2 \det\left(\mathbf{I}_M + \rho\mathbf{H}\mathbf{P}\mathbf{H}^\dagger\right) \tag{2.82}$$

$$= \log_2 \det\left(\mathbf{I}_M + \sum_{k=1}^{K} \rho_k \mathbf{h}_k \mathbf{h}_k^\dagger\right), \tag{2.83}$$

where $\rho_k = \rho \times p_k$ represents the total transmit power for user k. From here on in all MIMO capacity expressions provided assume the use of Gaussian symbols.

Using the eigendecomposition,

$$\mathbf{H}\mathbf{P}\mathbf{H}^\dagger = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\dagger, \tag{2.84}$$

where $\mathbf{U} \in \mathbb{C}^{M\times M}$ is a unitary matrix of eigenvectors and $\mathbf{\Lambda} \in \mathbb{C}^{M\times M}$ a diagonal matrix containing the ordered real-valued eigenvalues, λ_i, of which K are non-zero, the capacity may be written [208] as a sum of K SISO 'eigenchannels'

$$\mathcal{C}_{\text{SUM}}^{\text{UL}} = \sum_{i=1}^{K} \log_2\left(1 + \rho\lambda_i\right). \tag{2.85}$$

At high SNR (all $\rho\lambda_i \gg 1$) this is approximately

$$\mathcal{C}_{\text{SUM}}^{\text{UL}} \approx K \log_2 \rho + \sum_{i=1}^{K} \log_2 \lambda_i, \tag{2.86}$$

and the overall capacity increases with $K \log_2 \rho$, K times that achieved by the SISO or SIMO channel. The MU-MIMO system is therefore said to achieve a *spatial multiplexing gain* of K. Increasing the number of users (whilst ensuring $M \geq K$)[5] therefore allows significant throughput and sum spectral efficiency gains to be achieved, as illustrated in Figure 2.13.

The capacity is a function of the distribution of the channel eigenvalues. For a given received power,

$$P_R = \text{Tr}\left(\rho\mathbf{H}\mathbf{P}\mathbf{H}^\dagger\right) \tag{2.87}$$

$$= \rho \sum_{i=1}^{K} \lambda_i, \tag{2.88}$$

the maximum sum capacity is achieved when all channel eigenvalues are equal [178]. This occurs when the user channel vectors are mutually orthogonal, and the received power for each user is the same,

$$\mathbf{h}_k^T \mathbf{h}_j = 0, \quad \forall k, j \tag{2.89}$$

$$p_k \|\mathbf{h}_k\|^2 = p_j \|\mathbf{h}_j\|^2 \quad \forall k, j. \tag{2.90}$$

[5]Strictly, the condition $M \geq K$ is not necessary for reliable transmission of K user data streams, but *is* necessary to achieve a multiplexing gain of K. The MU-MIMO channel with $M < K$ is classed as *degraded*, and unavoidably results in the capacity of some user data streams being limited by inter-user interference.

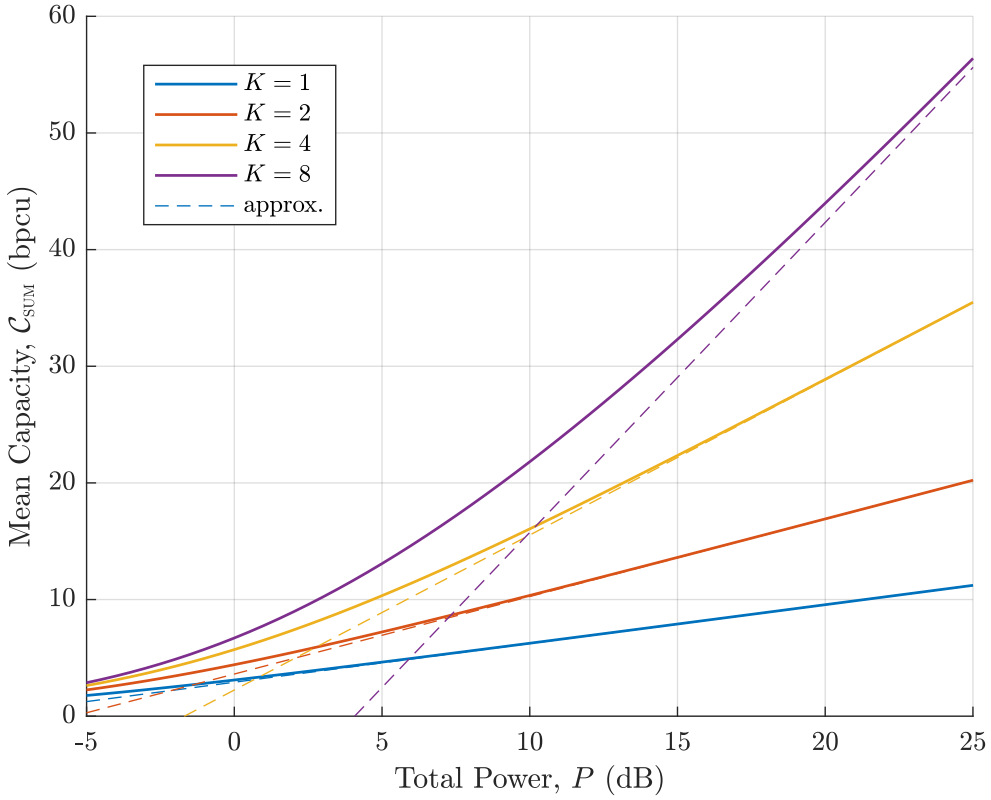

Figure 2.13: MU-MIMO sum capacity under i.i.d Rayleigh fading, $\rho_k = P/K$, $M = 8$

The worst sum capacity is achieved when only a single eigenvalue is non-zero (implying $\text{rank}(\mathbf{H}) = 1$). This occurs when all channel vectors are parallel,

$$\mathbf{h}_k \parallel \mathbf{h}_j \quad \forall k, j, \tag{2.91}$$

in which case spatial multiplexing cannot be achieved. In practical cellular propagation environments, the spatial separation of users will tend to ensure linear independence of the $\mathbf{h}_k$ and therefore that K distinct non-zero eigenvalues exist. A larger spread of these eigenvalues corresponds to a reduced MIMO channel capacity. A useful figure of merit is the channel condition number [212],

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}, \tag{2.92}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and minimum λ_i, respectively. Under i.i.d Rayleigh fading the channel condition improves (reduces) as M/K increases, as shown in Figure 2.14.

The sum capacity can be achieved by MMSE symbol detection with successive interference cancellation (MMSE-SIC, see Section 2.3.3), and defines the total achievable throughput of the MIMO channel [216]. The set of achievable user capacities then depends on the order in which the user symbols are detected and cancelled. In practice, for systems with many users the computational complexity of MMSE-SIC is prohibitive, and linear detection methods are often favoured, as discussed in Section 2.3.1.

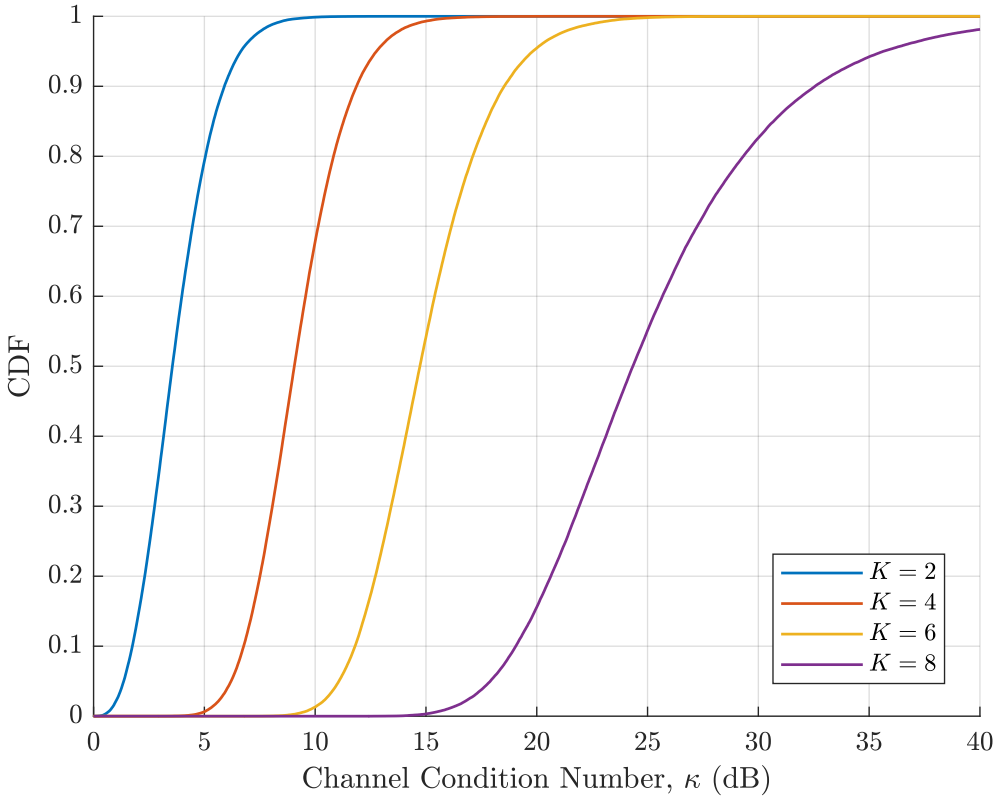

Figure 2.14: MIMO channel condition under i.i.d Rayleigh fading, $M = 8$.

MU-MIMO Downlink

An important and useful result from the theory of uplink-downlink duality is that the sum capacity of the downlink Gaussian MU-MIMO channel is identical to that of an equivalent uplink channel [218],

$$\mathcal{C}_{\text{SUM}}^{\text{DL}} = \log_2 \det \left(\mathbf{I}_M + \sum_{k=1}^{K} \rho_k \mathbf{h}_k \mathbf{h}_k^{\dagger} \right) \tag{2.93}$$

where ρ_k is the downlink power allocated to user data stream k. Under a total transmit power constraint, $P_T = \sum_k \rho_k$, sum capacity can be maximised by appropriate allocation of the ρ_k. Because of this duality, many conclusions about the behaviour of the uplink channel can be directly applied to the downlink channel.

A further result from uplink-downlink duality shows that the set of individual user capacities that can be achieved on the downlink, under a total transmit power constraint, are identical to those that can be achieved on the uplink dual channel using successive interference cancellation [218]. These downlink capacities are achieved by precoding using a combination of linear beamforming and 'dirty paper coding' (DPC) – wherein each user stream is coded to account for the interference generated by the other user data streams [200]. The complexity of implementing optimal DPC is often prohibitive in practical systems and linear beamforming methods, as discussed in the following section, are therefore often employed instead.

2.3 MU-MIMO Processing

The key signal processing tasks in a MU-MIMO system are the multi-user detection and precoding required to separate user data streams. Whilst the use of MMSE-SIC detection and DPC precoding achieve the maximum uplink and downlink MU-MIMO channel capacities, they are often impractical to implement when the number of users is large. The tasks of designing practical detection and precoding techniques have both achieved extensive research attention in the literature [238]. Here, particular attention is given to linear techniques that have complexity that scales well to large system, and are known to perform well when the channel matrix is well conditioned [178].

The discussion begins with linear detection methods, which are then extended to the downlink through uplink-downlink duality. A brief overview of alternative non-linear processing techniques, which can improve performance, is given at the end.

2.3.1 Linear Detection

With linear detection, each user symbol is estimated linearly before being individually decoded. The linear estimator takes the form

$$\hat{x}_k = \alpha_k \mathbf{w}_k^T \mathbf{y} \tag{2.94}$$

where $\mathbf{w}_k$ is the receive combining, or beamforming, vector for user k, and α_k a scalar that ensures correct output scaling by minimising $\mathbb{E}[|x_k - \hat{x}_k|^2]$. The resulting estimates contain a combination of the desired signal, the unwanted/interfering signals and noise

$$\hat{x}_k = \alpha_k \big(\sqrt{p_k} \mathbf{w}_k^T \mathbf{h}_k x_k + \sum_{j \neq k} \sqrt{p_j} \mathbf{w}_k^T \mathbf{h}_j x_j + \mathbf{w}_k^T \boldsymbol{\eta} \big) \tag{2.95}$$

$$= \text{Signal} + \text{Interference} + \text{Noise}, \tag{2.96}$$

and following the reasoning in Section 2.1.3 the capacity for user k under linear detection is therefore

$$\mathcal{C}_k^{\text{UL}} = \mathcal{I}(\hat{x}_k; x_k) \tag{2.97}$$

$$= \log_2 \left(1 + \text{SINR}_k^{\text{UL}} \right) \tag{2.98}$$

where

$$\text{SINR}_k^{\text{UL}} = \frac{\rho_k |\mathbf{w}_k^T \mathbf{h}_k|^2}{\sum_{j \neq k} \rho_j |\mathbf{w}_k^T \mathbf{h}_j|^2 + \|\mathbf{w}_k\|^2}. \tag{2.99}$$

The total capacity achievable under linear detection is the sum of the K user capacities

$$\mathcal{C}_{\text{LINEAR}}^{\text{UL}} = \sum_{k=1}^{K} \mathcal{C}_k^{\text{UL}}. \tag{2.100}$$

Note that the scaling α_k does not affect the estimate SINR or channel capacity, and therefore can be omitted from further analysis (but must be included when performing signal detection).

The linear estimation process can then be compactly written in matrix form

$$\hat{\mathbf{x}} = \mathbf{W}^{\mathrm{UL}}\mathbf{y} \tag{2.101}$$

where

$$\mathbf{W}^{\mathrm{UL}} = \begin{bmatrix} \mathbf{w}_1^T \\ \vdots \\ \mathbf{w}_K^T \end{bmatrix}. \tag{2.102}$$

Linear detection methods benefit from low complexity of detection – requiring only M scalar complex multiplications per user symbol. The computational complexity associated with calculating the detection vectors is higher, but since these can be re-used within a coherence block, the overall computational overheads are manageable for practical coherence block sizes.

Three widely studied linear detection methods are now outlined - the optimal linear minimum mean square estimation scheme, zero-forcing – which performs well at high SNR, and matched filtering – which performs well at low SNR.

Minimum Mean Square Error

The minimum mean square error (MMSE) estimator estimates each user symbol from $\mathbf{y}$ by treating the other interfering users as noise, and amongst all linear estimators achieves the highest user SINR. The MMSE receive combining vector is given by

$$\begin{aligned} \mathbf{w}_k^{(\mathrm{MMSE})} &= \arg\min_{\mathbf{w}} \quad \mathbb{E}\big[|\mathbf{w}^T\mathbf{y} - x_k|^2\big] \\ &= \mathbf{R}_{\mathrm{y}}^{-1}\mathbf{r}_{\mathrm{x}_k\mathrm{y}} \\ &= \Big(\mathbf{I}_M + \sum_{j\neq k}\rho_j\mathbf{h}_j\mathbf{h}_j^{\dagger}\Big)^{-1}\mathbf{h}_k^{*}\rho_k. \end{aligned} \tag{2.103}$$

Users capacities can be calculated from the SINRs given by (2.99). An alternative derivation exploits an underlying relationship between MMSE estimation and mutual information [212],

$$\begin{aligned} \mathcal{C}_k^{\mathrm{UL}} &= \mathcal{I}(\hat{x}_k; x_k) \\ &= \mathcal{I}(\mathbf{y}; x_k) \\ &= \mathcal{H}(\mathbf{y}) - \mathcal{H}(\mathbf{y}|x_k) \\ &= \mathcal{H}\Big(\sum_{j=1}^{K}\sqrt{p_j}\mathbf{h}_j x_j + \boldsymbol{\eta}\Big) - \mathcal{H}\Big(\sum_{j\neq k}\sqrt{p_j}\mathbf{h}_j x_j + \boldsymbol{\eta}\Big) \\ &= \log_2\big(1 + \rho_k\mathbf{h}_k^{\dagger}\Big(\mathbf{I}_M + \sum_{j\neq k}\rho_j\mathbf{h}_j\mathbf{h}_j^{\dagger}\Big)^{-1}\mathbf{h}_k\big). \end{aligned} \tag{2.104}$$

The second line in (2.104) follows since MMSE estimation, uniquely, is information lossless with respect to x_k. To reduce computational complexity, the K MMSE estimation vectors can be calculated simultaneously using

$$\mathbf{W}^{(\mathrm{MMSE})} = \mathbf{P}^{-1/2}\Big(\mathbf{H}^{\dagger}\mathbf{H} + \rho^{-1}\mathbf{P}^{-1}\Big)^{-1}\mathbf{H}^{\dagger}, \tag{2.105}$$

which requires a single $K \times K$ matrix inversion rather than K separate $M \times M$ matrix inversions. Similarly, the SINR can be written [138]

$$\mathrm{SINR}_k^{(\mathrm{MMSE})} = \frac{1}{\left[\left(\mathbf{I}_K + \rho \mathbf{P}^{1/2}\mathbf{H}^\dagger \mathbf{H}\mathbf{P}^{1/2}\right)^{-1}\right]_{k,k}} - 1. \tag{2.106}$$

Zero Forcing

The zero forcing (ZF) estimator minimises the estimate noise whilst eliminating *all* inter-user interference, i.e.

$$\begin{aligned} \mathbf{w}_k^{(\mathrm{ZF})} = \ & \arg\min_{\mathbf{w}} \ \|\mathbf{w}\|^2 \\ & \text{subject to} \ \ \mathbf{w}^T \mathbf{h}_j = \delta_{j,k}, \end{aligned} \tag{2.107}$$

where

$$\delta_{j,k} = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases} \tag{2.108}$$

The ZF vectors are most concisely expressed in matrix form as the Moore-Penrose pseudo-inverse

$$\mathbf{W}^{(\mathrm{ZF})} = \left(\mathbf{H}^\dagger \mathbf{H}\right)^{-1} \mathbf{H}^\dagger. \tag{2.109}$$

This is also known as channel inversion, since

$$\mathbf{W}^{(\mathrm{ZF})} \mathbf{H} = \mathbf{I}_K. \tag{2.110}$$

The symbol estimate produced by ZF contains only receiver noise

$$\hat{x}_k = x_k + \mathbf{w}_k^{(\mathrm{ZF})T} \boldsymbol{\eta}, \tag{2.111}$$

and has SINR

$$\mathrm{SINR}_k^{(\mathrm{ZF})} = \frac{\rho_k}{\|\mathbf{w}_k^{(\mathrm{ZF})}\|^2}. \tag{2.112}$$

The term $\|\mathbf{w}_k^{(\mathrm{ZF})}\|^2$ is sometimes called the noise amplification factor,

$$\|\mathbf{w}_k^{(\mathrm{ZF})}\|^2 = \left[\left(\mathbf{H}^\dagger \mathbf{H}\right)^{-1}\right]_{k,k} \tag{2.113}$$

and its value reflects how 'difficult' it is to eliminate inter-user interference from a given user stream.

Matched Filter

The matched filter from Section 2.2.1,

$$\mathbf{w}_k^{(\mathrm{MF})} = \mathbf{h}_k^*, \tag{2.114}$$

or

$$\mathbf{W}^{(\mathrm{MF})} = \mathbf{H}^\dagger, \tag{2.115}$$

maximises signal to noise ratio (minimises noise amplification), without accounting for inter-user interference. The SINR of stream k under matched filtering is

$$\mathrm{SINR}_k^{(\mathrm{MF})} = \frac{\rho_k \|\mathbf{h}_k\|^2}{\sum_{j \neq k} \rho_j \frac{|\mathbf{h}_k^\dagger \mathbf{h}_j|^2}{\|\mathbf{h}_k\|^2} + 1}. \tag{2.116}$$

Since MF detection does not attempt to reduce inter-user interference it generally performs poorly in multi-user scenarios. However, at low SNR, where noise dominates over interference, MF detection is close to optimal.

Performance Comparison

The MMSE detector is the optimal linear detector in that it optimally balances inter-user interference against noise amplification. Despite this, it does not generally fully achieve the sum capacity of the uplink channel as given in (2.80), since – unlike the MMSE-SIC detector – estimation and detection is performed independently for each user symbol.

Using (2.106) and a method similar to [87] it is straightforward to show that the individual user capacities under MMSE detection are bounded according to the channel eigenvalues

$$\log_2(1 + \rho\lambda_{\min}) \leq \mathcal{C}_k^{(\mathrm{MMSE})} \leq \log_2(1 + \rho\lambda_{\max}). \tag{2.117}$$

When the channel condition number, κ, approaches 1 these bounds converge and MMSE detection comes close to achieving the full channel sum capacity. A second, tighter, upper bound on per-user capacity is the single-user bound,

$$\mathcal{C}_k^{(\mathrm{MMSE})} \leq \log_2(1 + \rho_k \|\mathbf{h}_k\|^2), \tag{2.118}$$

which is the capacity of the SIMO channel where only user k transmits.

Comparing (2.109) to (2.105) it can be seen that for high ρ

$$\mathbf{W}^{(\mathrm{ZF})} \approx \mathbf{P}^{1/2}\mathbf{W}^{(\mathrm{MMSE})}, \tag{2.119}$$

and ZF detection and MMSE detection become equivalent (up to a scaling factor). On the other hand, comparing (2.115) to (2.105) it can be seen that for low ρ,

$$\rho\mathbf{W}^{(\mathrm{MF})} \approx \mathbf{P}^{-1/2}\mathbf{W}^{(\mathrm{MMSE})}, \tag{2.120}$$

MF detection is approximately optimal.

Under i.i.d Rayleigh fading, both MMSE and ZF detection give a diversity gain of $(M-K+1)$ per user stream [88], compared to M for the SIMO channel. This loss in diversity is due to a sacrifice of degrees of freedom when reducing/eliminating inter-user interference. Under linear detection an excess of BS antennas are therefore required ($M > K$) to achieve a multi-antenna diversity gain for the receivers. The trade-off between multiplexing and diversity in multiple antenna channels is an important aspect of algorithm design and analysis [211], but a detailed

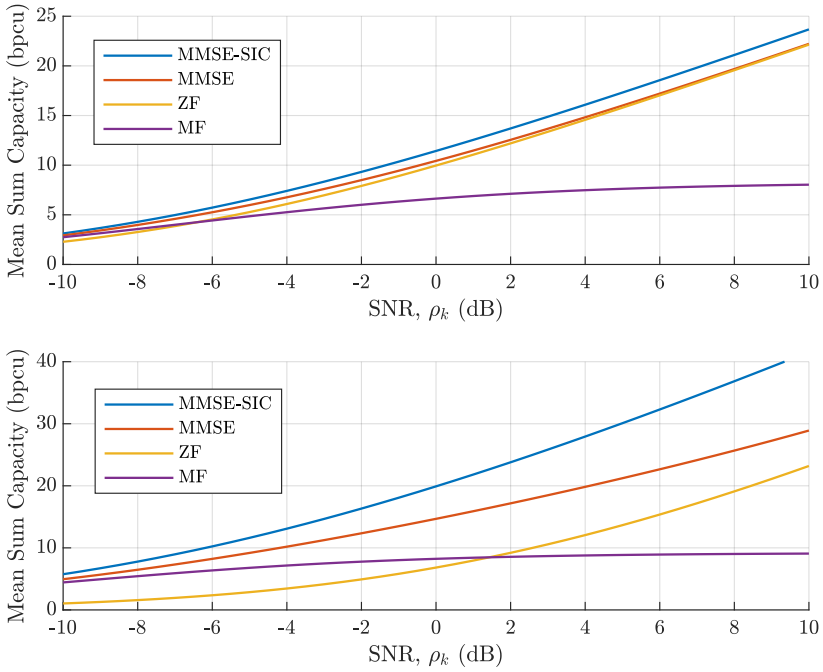

Figure 2.15: MIMO capacity under linear detection with i.i.d Rayleigh fading, top: $M = 8, K = 4$, bottom: $M = 8, K = 8$

study of it is outside the scope of this work.

Figure 2.15 compares the total capacity achieved under the different detection strategies. When $M = 8, K = 4$ linear methods come close to achieving the full capacity, whilst for $M = 8, K = 8$ the poor channel condition means that they perform poorly compared to optimal MMSE-SIC detection.

2.3.2 Linear Precoding

Under linear precoding, the transmit signal is a linear combination of the user symbols, where each symbol is weighted with a beamforming vector, $\mathbf{w}_k$,

$$\mathbf{x} = \sum_{k=1}^{K} \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|} \sqrt{p_k} s_k \tag{2.121}$$

with normalisation to ensure that $P_T = \mathbb{E}[\|\mathbf{x}\|^2] = \sum_{k=1}^{K} \rho_k$. The received signal at user k again consists of signal, interference and noise,

$$y_k = \frac{\mathbf{h}_k^T \mathbf{w}_k}{\|\mathbf{w}_k\|} \sqrt{p_k} s_k + \sum_{j \neq k} \frac{\mathbf{h}_k^T \mathbf{w}_j}{\|\mathbf{w}_j\|} \sqrt{p_j} s_j + \eta_k, \tag{2.122}$$

and therefore has capacity

$$\mathcal{C}_k^{\mathrm{DL}} = \log_2 \left(1 + \mathrm{SINR}_k^{\mathrm{DL}}\right), \tag{2.123}$$

where

$$\mathrm{SINR}_k^{\mathrm{DL}} = \frac{\rho_k |\mathbf{h}_k^T \mathbf{w}_k|^2}{\sum_{j \neq k} \rho_j |\mathbf{h}_k^T \mathbf{w}_j|^2 \frac{\|\mathbf{w}_k\|^2}{\|\mathbf{w}_j\|^2} + \|\mathbf{w}_k\|^2}. \tag{2.124}$$

The precoding can be written in matrix form

$$\mathbf{x} = \mathbf{W}^{\mathrm{DL}} \mathbf{P}^{1/2} \mathbf{s}, \tag{2.125}$$

where

$$\mathbf{W}^{\mathrm{DL}} = \begin{bmatrix} \mathbf{w}_1 & \dots & \mathbf{w}_K \end{bmatrix}, \tag{2.126}$$

and

$$\mathbf{P} = \mathrm{diag}\Big(\frac{p_k}{\|\mathbf{w}_k\|^2} \Big). \tag{2.127}$$

At the users, a scaling factor is applied for optimal scaling before detection [102],

$$\hat{s}_k = \alpha_k y_k. \tag{2.128}$$

Uplink-Downlink Duality

An important and remarkable result from uplink-duality [218] says that under a total transmit power, $\sum_{k=1}^{K} \rho_k^{\mathrm{DL}} = \sum_{j=1}^{K} \rho_j^{\mathrm{UL}}$, the same user SINRs can be achieved on both the uplink and downlink,

$$\mathrm{SINR}_k^{\mathrm{DL}} = \mathrm{SINR}_k^{\mathrm{UL}} \quad \forall k, \tag{2.129}$$

using the same beamforming vectors

$$\mathbf{w}_k^{\mathrm{DL}} = \mathbf{w}_k^{\mathrm{UL}} \quad \forall k, \tag{2.130}$$

and a different set of power allocations

$$\rho_k^{\mathrm{DL}} \neq \rho_k^{\mathrm{UL}} \tag{2.131}$$

Any linear precoding design can therefore be recast and analysed as a dual *virtual* linear detection problem with

$$\mathbf{W}^{\mathrm{DL}} = \big(\mathbf{W}^{\mathrm{UL}} \big)^T. \tag{2.132}$$

For reciprocal channels, this may also enable receive beamforming vectors to be re-used on both uplink and downlink, although since an uplink system will typically operate under per-user power constraints, $\rho_k \leq P$, whilst the downlink may operate under a total power constraint, $\sum_k \rho_k \leq P_T$, the optimal (LMMSE) beamforming vectors will differ between the uplink and downlink.

Simplified Uplink Channel & Downlink Precoding Model

For ease of analysis, it is convenient to remove the uplink user power scaling matrix by absorbing it into the channel matrix,

$$\mathbf{y}^{\mathrm{UL}} = \mathbf{H}^{\mathrm{UL}}\mathbf{x}^{\mathrm{UL}} + \boldsymbol{\eta} \tag{2.133}$$

where

$$\mathbf{H}^{\mathrm{UL}} = \left[\sqrt{p_1}\mathbf{h}_1^{\mathrm{UL}} \quad \dots \quad \sqrt{p_K}\mathbf{h}_K^{\mathrm{UL}}\right]. \tag{2.134}$$

Similarly, it is convenient to absorb the power control factors into the beamforming vectors, $\mathbf{w}_k$, reducing the downlink precoding model to

$$\mathbf{x}^{\mathrm{DL}} = \mathbf{W}^{\mathrm{DL}}\mathbf{s}, \tag{2.135}$$

with SINR,

$$\mathrm{SINR}_k^{\mathrm{DL}} = \frac{\rho_k|\mathbf{h}_k^T\mathbf{w}_k|^2}{\sum_{j\neq k}\rho_j|\mathbf{h}_k^T\mathbf{w}_j|^2 + 1}. \tag{2.136}$$

This model simplifies notation, and is used throughout this thesis.

2.3.3 Non-Linear Methods

The optimal MMSE-SIC and DPC methods, and other non-linear methods that approximate their performance, can offer potential performance improvements over linear methods.

MMSE-SIC & V-BLAST

The capacity-achieving MMSE-SIC detector has a structure that follows directly from the chain rule of mutual information [216],

$$\begin{aligned}\mathcal{C}_{\mathrm{SUM}}^{\mathrm{UL}} &= \mathcal{I}(\mathbf{x};\mathbf{y}) && (2.137)\\ &= \mathcal{I}(x_1;\mathbf{y}) + \mathcal{I}(x_2;\mathbf{y}|x_1) && (2.138)\\ &\quad + \mathcal{I}(x_3;\mathbf{y}|x_1,x_2) + \sum_{k=4}^{K}\mathcal{I}(x_k;\mathbf{y}|x_1,\dots,x_{k-1}).\end{aligned}$$

The first term in (2.137) is simply the mutual information captured by the MMSE estimate of x_1 from $\mathbf{y}$. The second term is then the mutual information captured by the MMSE estimate of x_2 from $\mathbf{y}$ when x_1 is known, the third the MMSE estimate with x_1 and x_2 known, and so on. Optimal detection can therefore be achieved by first finding the MMSE estimate of x_1 and decoding the first data stream. If x_1 is encoded at a rate less than

$$\mathcal{C}_1^{\mathrm{MMSE-SIC}} = \log_2\left(1 + \rho_1\mathbf{h}_1^{\dagger}\Big(\mathbf{I}_M + \sum_{k=2}^{K}\rho_k\mathbf{h}_k\mathbf{h}_k^{\dagger}\Big)^{-1}\mathbf{h}_1\right), \tag{2.139}$$

it can be decoded error free. The interference generated by x_1 is then subtracted from the received signal, and x_2 is estimated. Since this estimate contains no interference from user 1,

x_2 may encoded at a higher rate than under MMSE detection. The capacity of user k is

$$\mathcal{C}_k^{\text{MMSE-SIC}} = \log_2\left(1+\rho_k \mathbf{h}_k^\dagger \Big(\mathbf{I}_M + \sum_{j=k+1}^{K} \rho_j \mathbf{h}_j \mathbf{h}_j^\dagger\Big)^{-1} \mathbf{h}_k\right) \tag{2.140}$$

$$> \log_2\left(1+\rho_k \mathbf{h}_k^\dagger \Big(\mathbf{I}_M + \sum_{j\neq k}^{K} \rho_j \mathbf{h}_j \mathbf{h}_j^\dagger\Big)^{-1} \mathbf{h}_k\right), \tag{2.141}$$

with symbol estimate k given by

$$\hat{x}_k = \mathbf{w}_k^T \mathbf{y}_k \tag{2.142}$$

where

$$\mathbf{w}_k^{(\text{SIC})} = \Big(\mathbf{I}_M + \sum_{j=k+1}^{K} \rho_j \mathbf{h}_j \mathbf{h}_j^\dagger\Big)^{-1} \mathbf{h}_k \rho_k^{-1} \tag{2.143}$$

and

$$\mathbf{y}_k = \mathbf{y} - \sum_{j=1}^{k-1} \sqrt{p_j}\mathbf{h}_j \mathbf{h}_j^\dagger x_j. \tag{2.144}$$

Clearly the numbering of the users, which is arbitrary, determines the order of cancellation and therefore the user capacities. The $K!$ potential user orderings and power allocations ρ_k define this region of achievable user transmission rates.

The MMSE-SIC detector relies on each user stream being fully decoded before subsequent streams can be detected. Since long coding blocks are required for optimal performance, this can introduce latency and increase data buffering requirements.

A practical alternative is to make hard symbol decoding decisions based on individual symbol estimates, without applying error correction decoding, and use these symbols to cancel the interference. This is the basis of the revolutionary V-BLAST (Vertical Bell Laboratories Layered Space-Time) experimental system [228], a key early demonstrator of the potential of MIMO systems.

The V-BLAST scheme is a variant of the decision feedback equaliser used in time domain equalisation systems [68], and can suffer from the same error-propagation issues, where errors in the hard decisions propagate and cause errors in subsequently decoded streams. At high SNR the performance of V-BLAST is therefore limited by the first user symbol that is detected, which has diversity order $d = M-K+1$ under i.i.d Rayleigh fading (the same as linear MMSE). Error-propagation can be minimised by optimising the cancellation order, so that symbols with lower probability of error (higher SINR) are cancelled first [125]. Whilst the use of optimal cancellation ordering does not improve the diversity order, it can significantly improve bit error rate by providing an effective SINR gain [240].

Maximum Likelihood Estimation & Sphere Decoding

In practical wireless systems using symbols drawn from discrete lattice alphabets, such as QAM, symbol estimation is performed over a discrete search space, and optimal detection becomes more challenging [85].

For a single channel use, the maximum likelihood symbol estimate is found by minimising,

over the set of discrete candidate symbol vectors, $\mathbf{x} \in \mathcal{A}$, the Euclidean distance between the received signal and the candidate transmitted vector

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathcal{A}}{\arg\min} \quad \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2. \tag{2.145}$$

For coded transmission over multiple channel uses this can be modified to produce soft probability outputs for use in error correction decoding. Under i.i.d Rayleigh fading, ML detection achieves the maximum diversity order of $d = M$ [215] – which can represent a significant improvement over linear MMSE detection when the number of receive antennas is comparable to the number of users. Unfortunately, for user constellation size 2^b there are $|\mathcal{A}| = 2^{bK}$ candidate solutions to test, and the computational complexity of ML estimation increases exponentially with the number of users.

Sphere decoding decreases the complexity of ML estimation by limiting the search to those candidate symbols that fall within a sphere of certain radius around the received signal, $\mathbf{x} \in \tilde{\mathcal{A}}$,

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\tilde{\mathcal{A}}}{\arg\min} \quad \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2. \tag{2.146}$$

This can reduce the expected complexity to be polynomial, and often roughly cubic in the codeword size [85].

Figure 2.16 compares the bit error rate achieved under linear (MMSE) and non-linear (MMSE V-BLAST, sphere decoding) detection in two i.i.d Rayleigh channels. The sphere decoder is implemented using Matlab's Communications Toolbox [209]. In the first case ($M = 8, K = 2$), there is a significant excess of receive antennas and the benefits from using non-linear detection are small. Increasing the number of users to $K = 6$ degrades the performance of MMSE detection, and non-linear detection becomes more beneficial. The use of optimised cancellation ordering greatly improves the performance of V-BLAST, with sphere decoding giving the best performance.

Dirty Paper Coding & Non-Linear Precoding

On the downlink, the optimal precoding method is based on the landmark dirty paper coding result from Costas, which showed that the capacity of a channel with receiver interference known to the transmitter is identical to the channel capacity without interference [45]. Successive linear beamforming and dirty paper coding of user symbols, where preceding and subsequent user symbols are treated as known and unknown interference respectively, provides a theoretical precoder structure capable of achieving the sum capacity of the MU-MIMO downlink channel, and is a direct downlink dual of MMSE-SIC detection [218].

Practical sub-optimal DPC-style precoders based on linear beamforming with Tomlinson-Harashima precoding have been applied to MU-MIMO systems with discrete signal constellations and can offer performance gains over linear precoding [225], and linear beamforming with a vector peturbation of the user symbols can achieve near-optimal performance whilst only requiring a simple additional modulo operation at the receivers [89].

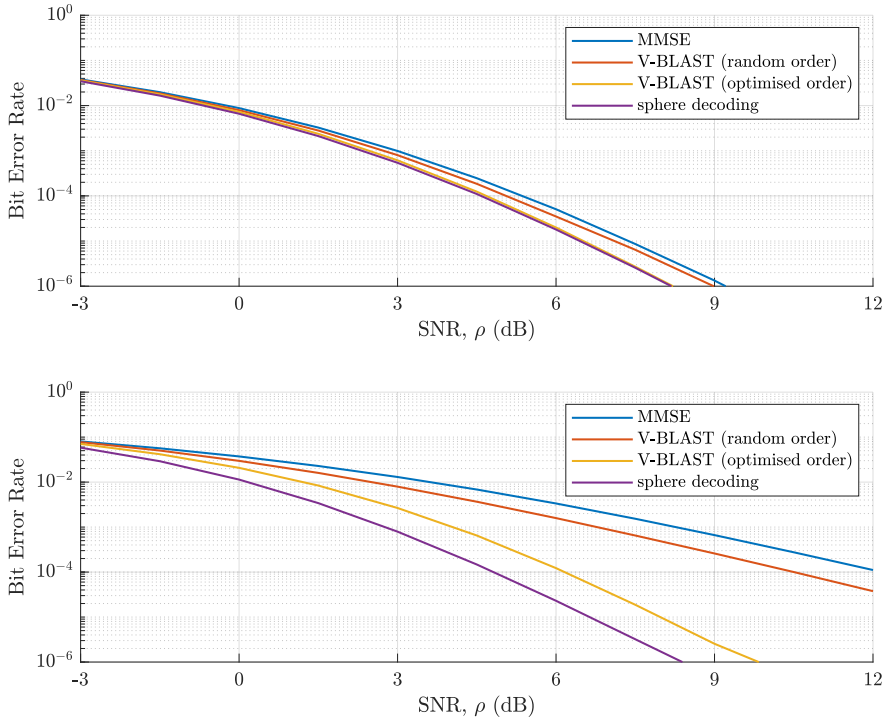

Figure 2.16: Bit error rate of QPSK modulation with MMSE, MMSE V-BLAST (random & optimised cancellation orders) and sphere decoding in i.i.d Rayleigh channels, top: $M = 8, K = 2$, bottom: $M = 8, K = 6$

2.4 MU-MIMO Channels

As seen above, the performance of an MU-MIMO system under fading depends on the statistical distributions of the user channel vectors. For the purpose of theoretical analysis the entries of the channel vectors are generally modelled as independent variables following a Rayleigh distribution, and assumed to be perfectly known to the user and base station. This allows for easy comparison of different systems and algorithms, and often enables insightful closed form expressions to be obtained [144] – increasingly through the use of random matrix theory [46]. However, in accurately modelling practical MIMO systems these assumptions often fail on both counts – the channel entries being neither independently Rayleigh fading nor perfectly known.

This section discusses the importance of accounting for correlated fading in evaluating MIMO system performance, and outlines simple ray-based and correlated Rayleigh fading models for doing so. Channel estimation is then discussed, and a method for analysing the impact of imperfect CSI on uplink and downlink capacity introduced. Finally, power control methods for combatting large scale fading are outlined.

2.4.1 Correlated Fading Channel Models

The use of i.i.d Rayleigh fading is based on the reasoning provided in Section 2.1.4, with the additional assumption that the channel coefficients for each antenna fade independently. Physically, this implies that the signals transmitted by the user arrive at the base station uniformly from all directions [15], and assumes the use of physically unrealisable isotropic antennas [178]. This is a reasonable approximation for indoor environments with rich multipath scattering, but is not valid in larger outdoor settings where propagation is often strongly directional due to the physical layout of the propagation environment. Measurement results show that real life capacities can differ significantly to those predicted by independent fading [38].

When fading is instead *correlated* across antennas, system performance is affected. For example, there is an increase in the likelihood that when one antenna experiences a deep fade others also do and hence the effective diversity gain is reduced [171]. Similarly, eliminating inter-user interference becomes more challenging for closely spaced users when cell geometry means that their channels are highly correlated [106]. The evaluation of communications systems and algorithm performance under correlated fading is thus necessary to gain a full picture.

Research into MIMO channel models that more accurately capture the spatial characteristics of both indoor and outdoor wireless environments scenarios has been extensive – see, for example, [43] for a thorough survey. A detailed discussion or comparison of these models is well beyond the scope and purpose of this thesis. Instead, a simple ray-based channel model is outlined that captures spatial characteristics of the channel. This is then extended to a correlated Rayleigh fading model. Using channels captured with a ray tracing tool, these channel models are used later to evaluate algorithm performance.

Deterministic Ray-based Channel

The deterministic path based propagation model in Section 2.1.4 is readily extended to account for spatial dependencies in the channel by assuming that user transmissions arrive at the BS

antenna array with a plane wavefront [15]. This effectively assumes that the same propagation paths exist between each user and BS antenna, with the spacing of the BS antennas meaning the time delays of the paths vary between antennas. This is a reasonable modelling assumption whenever the user and reflective clusters in the environment are far enough away from the BS that they lie in the far-field region of the array. This is illustrated in Figure 2.17 for a 1-dimensional array.

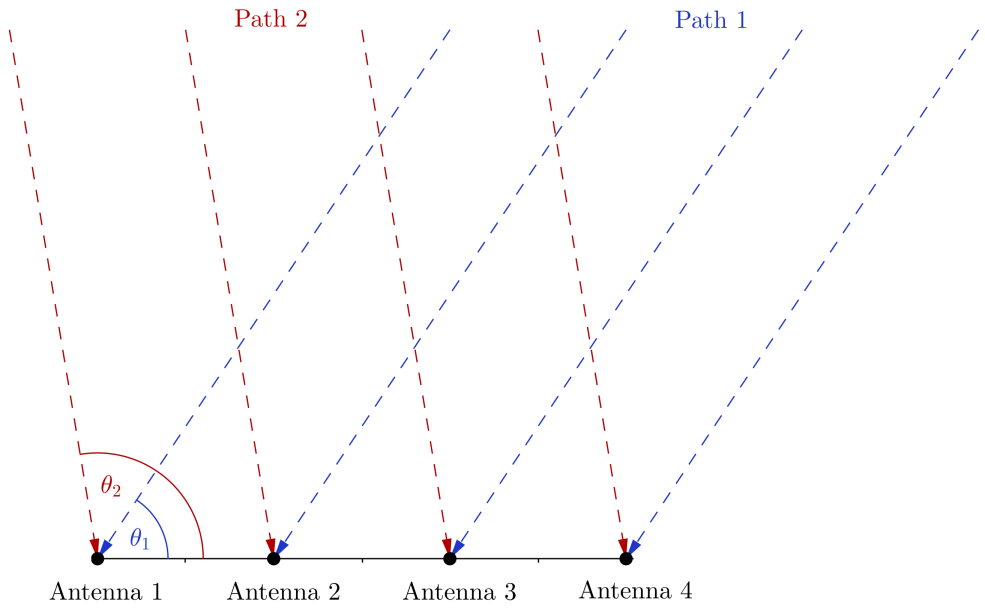

Figure 2.17: Two planewaves incident on four-element 1-dimensional array.

The channel vector for user k is then given by

$$\mathbf{h}_k = \sum_i \alpha_{k,i} \mathbf{a}(\theta_{k,i}, \phi_{k,i}) \tag{2.147}$$

where $\alpha_{k,i}$ is a complex coefficient describing the propagation path to the reference BS antenna, $\theta_{k,i}$ and $\phi_{k,i}$ are the azimuth and elevation angles of arrival of the path, and $\mathbf{a}(\theta_{k,i}, \phi_{k,i})$ is a steering vector of the form,

$$\mathbf{a}(\theta_{k,i}, \theta_{k,i}) = \left[e^{j\Omega_1(\theta_{k,i},\theta_{k,i})} \quad \dots \quad e^{j\Omega_M(\theta_{k,i},\theta_{k,i})} \right]^T \tag{2.148}$$

where $\Omega_m(\theta_k, \phi_k)$ is a phase shift at antenna m with respect to the path to the reference antenna, caused by the differing time delays [15]. Under this model, when a single strongly dominant path exists between a user and the array the channel coefficients for each antenna vary only by a phase shift.

Reciprocity of electromagnetic propagation means that this model is equally valid on the downlink where transmissions are from the BS to the user.

Correlated Rayleigh Fading Channel

As in Section 2.1.4, the deterministic model above can be extended to a random fading model by treating the complex coefficients $\alpha_{k,i}$ as random variables. For multipath components with distinct angles of arrival, the $\alpha_{k,i}$ follow a circular uniform distribution, and can be modelled as varying independently. In propagation environments with clusters of components arriving from the same angle of arrival, it can be more appropriate to treat the $\alpha_{k,i}$ as independent Rayleigh fading variables [15].

Assuming all of the paths (or path clusters) have similar power, by the central limit theorem the channel can modelled by correlated Rayleigh fading,

$$\mathbf{h}_k \sim \mathcal{CN}\big(0, \mathbf{R}_k\big) \tag{2.149}$$

where $\mathbf{R}_k$ is the channel covariance for user k,

$$\mathbf{R}_k = \mathbb{E}\big[\mathbf{h}_k \mathbf{h}_k^\dagger\big] \tag{2.150}$$

$$= \sum_i \mathbb{E}\big[|\alpha_{k,i}|^2\big] \mathbf{a}(\theta_{k,i}, \phi_{k,i}) \mathbf{a}(\theta_{k,i}, \phi_{k,i})^\dagger. \tag{2.151}$$

This is a generalised Rayleigh fading model for which independent fading is a specific case ($\mathbf{R}_k = \mathbf{I}_M$), requiring at least M multipath components to exist. This model assumes that the channels are wide-sense stationary – the multipath structure stays the same over time. Thus it is valid if the users only undergo small displacements, but not when the users undergo large displacements such that the multipath strengths and angles of arrivals change.

The diagonal elements of the covariance matrix represent the large scale fading (average channel strength)

$$\big[\mathbf{R}_k\big]_{m,m} = \beta_k \tag{2.152}$$

$$= \sum_i \mathbb{E}\big[|\alpha_{k,i}|^2\big]. \tag{2.153}$$

Channel covariance matrices can be constructed by generating random arrival paths according to a statistical angle of arrival model, and normalising according to a chosen pathloss model. Accurate modelling of the angle of arrival models is very involved [145], but a model based on a Laplacian distributed angular spread in the azimuth direction [167] and delta distributed elevation (i.e no spread) has been shown to be reasonable, with azimuth spreads of $\sim 10^\circ$ (standard deviation) representative of urban propagation environments [144].

2.4.2 Channel Estimation

All of the processing methods outlined in Section 2.3 require explicit knowledge of the user channel vectors. In practice this channel state information (CSI) cannot be known a priori and must be estimated from measurements of signals received through the channel.

A popular method of estimating the channel is through the transmission of pilot sequences, known to both transmitter and receiver. This section outlines a basic LMMSE channel estima-

tion scheme that uses uplink pilots and prior knowledge of the channel statistics to estimate the channel.

Uplink Pilot-based LMMSE Channel Estimation

Under uplink channel estimation each user transmits a pilot sequence of N_s predefined symbols, $\boldsymbol{\varphi}_k \in \mathbb{C}^{N_s}$ within a single coherence block, where $\|\boldsymbol{\varphi}_k\|^2 = N_s$. Assuming pilot SNR ρ_{CSI}, the received sequence of symbols, $\mathbf{Y} \in \mathbb{C}^{M \times N_s}$, is given by

$$\mathbf{Y} = \sqrt{\rho_{\text{CSI}}} \sum_{k=1}^{K} \sqrt{p_k} \mathbf{h}_k \boldsymbol{\varphi}_k^{\dagger} + \mathbf{N} \tag{2.154}$$

with receiver noise $[\mathbf{N}]_{i,j} \sim \mathcal{CN}(0,1)$. At the receiver, a statistic $\mathbf{t}_k$ is then formed for each user by correlating the received symbols with the appropriate pilot sequence,

$$\begin{aligned} \mathbf{t}_k &= \frac{1}{\sqrt{N_s}} \mathbf{Y} \boldsymbol{\varphi}_k \\ &= \sqrt{\rho_{\text{CSI}}} \sqrt{p_k} \mathbf{h}_k + \sqrt{\frac{\rho_{\text{CSI}}}{N_s}} \sum_{j \neq k} \sqrt{p_j} \mathbf{h}_j \boldsymbol{\varphi}_j^{\dagger} \boldsymbol{\varphi}_k + \boldsymbol{\eta}. \end{aligned} \tag{2.155}$$

For the case of orthogonal pilots ($\boldsymbol{\varphi}_j^{\dagger} \boldsymbol{\varphi}_k = N_s \delta_{j,k}$) this reduces to

$$\mathbf{t}_k = \sqrt{\rho_{\text{CSI}}} \sqrt{p_k} \mathbf{h}_k + \boldsymbol{\eta}. \tag{2.156}$$

The use of orthogonal pilots requires $N_s \geq K$, and thus increasing the number of users increases the transmission overhead associated with channel estimation.

The channel vector can then be estimated from the statistic. Linear minimum mean square error (LMMSE) channel estimation requires knowledge of only the first and second order statistics of the user channel vectors, and represents an attractive means of estimating the channel when these statistics are available. For the case of correlated Rayleigh fading it is also the optimal estimation method. Assuming $\mathbb{E}\big[\mathbf{h}_k\big] = \mathbf{0}$, the LMMSE estimate is given by

$$\hat{\mathbf{h}}_k = \frac{1}{\sqrt{\rho_{\text{CSI}}} \sqrt{p_k}} \mathbf{R}_k \Big(\mathbf{R}_k + \frac{1}{\rho_{\text{CSI}} p_k} \mathbf{I}_M \Big)^{-1} \mathbf{t}_k. \tag{2.157}$$

This results in an estimate of the form

$$\hat{\mathbf{h}}_k = \mathbf{h}_k - \mathbf{e}_k, \tag{2.158}$$

where $\mathbf{e}_k$ is the estimation error, which has covariance

$$\begin{aligned} \mathbf{C}_k &= \mathbb{E}\big[\mathbf{e}_k \mathbf{e}_k^{\dagger}\big] \\ &= \big(\mathbf{R}_k^{-1} + \rho_{\text{CSI}} p_k \mathbf{I}_M \big)^{-1}, \end{aligned} \tag{2.159}$$

and, by the orthogonality principle, is uncorrelated with the channel estimate

$$\mathbb{E}\big[\hat{\mathbf{h}}_k \mathbf{e}_k^{\dagger}\big] = \mathbf{0}. \tag{2.160}$$

LMMSE channel estimation requires knowledge of the channel covariance matrices at the base station, and in practice these must themselves be estimated – exploiting the fact that the channel covariance matrices stays fixed across multiple coherence blocks when the multipath structure stays the same [15]. Alternatively, channel estimation methods that do not require channel statistics may be used, such as least squares estimation,

$$\hat{\mathbf{h}}_k = \frac{\mathbf{t}_k}{\sqrt{\rho_{\text{CSI}}}\sqrt{p_k}}, \tag{2.161}$$

but these result in more channel estimation error.

In principle, a similar scheme can be employed using downlink pilots to estimate the downlink channels at the users. However, since this downlink CSI is required at the BS, these channel estimates must then be fed back to the BS. Furthermore, the required length of the downlink pilot sequences grows with the number of BS antennas, and can become prohibitive for MIMO systems operating with a large number of BS antennas. Research attention has therefore been given to CSI quantization and approximation schemes to reduce the overheads associated with downlink channel estimation, e.g. [40].

When uplink and downlink transmission occurs within the same frequency channel – as in time division duplex (TDD) systems – CSI feedback can be avoided by exploiting channel reciprocity. Using the fact that the uplink and downlink propagation channel are identical within a coherence block, channels measured on the uplink can then be used for downlink precoding (after appropriate calibration to account for hardware different in the transmit and receive chains affecting the channel response), reducing overheads.

The treatment of imperfect CSI throughout this thesis assumes that LMMSE channel estimation is used, with channel reciprocity exploited for downlink CSI.

Uplink Channel Capacity with Imperfect CSI

The channel capacities given in Sections 2.2.3 & 2.3 only apply in the case where perfect channel state information is available. If the channel estimates used for data detection and precoding have errors then performance will be reduced. In general, full characterisation of the MIMO channel capacity under imperfect CSI remains an open problem. However, a useful and achievable lower bound is obtained for correlated Rayleigh fading channels, $\mathbf{h}_k \sim \mathcal{CN}(0, \mathbf{R}_k)$, by assuming that the signal through the unknown channel component, $\mathbf{e}_k$, is treated as noise [84].

On the uplink, the received signal may be decomposed into parts received through the known and unknown channel

$$\mathbf{y} = \sum_{k=1}^{K} \hat{\mathbf{h}}_k \sqrt{p_k} x_k + \sum_{k=1}^{K} \mathbf{e}_k \sqrt{p_k} x_k + \boldsymbol{\eta} \tag{2.162}$$

$$= \hat{\mathbf{H}} \mathbf{P}^{1/2} \mathbf{x} + \mathbf{E} \mathbf{P}^{1/2} \mathbf{x} + \boldsymbol{\eta}. \tag{2.163}$$

The component through the unknown channel and the receiver noise may then be combined

into an equivalent noise term, $\boldsymbol{\nu} = \sum_{k=1}^{K} \mathbf{e}_k \sqrt{p_k} x_k + \boldsymbol{\eta}$, giving

$$\mathbf{y} = \sum_{k=1}^{K} \sqrt{p_k} \hat{\mathbf{h}}_k x_k + \boldsymbol{\nu}. \tag{2.164}$$

Since the realisation of $\mathbf{e}_k$ is unknown, for a given channel estimate the statistics of this equivalent noise are unknown. However, over possible channel estimates $\mathbf{e}_k$ can be treated as a random variable with variance,

$$\mathbb{E}_{\mathbf{E}}\big[\boldsymbol{\nu}\boldsymbol{\nu}^{\dagger}\big] = \boldsymbol{\Omega} \tag{2.165}$$

$$= \mathbf{I}_M + \sum_{k=1}^{K} \rho_k \mathbf{C}_k, \tag{2.166}$$

where $\mathbb{E}_{\mathbf{E}}[.]$ is the expectation with respect to both transmit symbols and channel estimation error realisations. From the orthogonality principle, (2.160), this equivalent noise is uncorrelated with the received signal. With the receiver treating the channel estimation error as equivalent noise, it is shown in [84] that for Rayleigh fading channels the expected capacity is given by

$$\overline{\mathcal{C}}_{\text{SUM}} = \mathbb{E}_{\mathbf{E}}\big[\log_2 \det\big(\mathbf{I}_K + \rho \mathbf{P}^{1/2} \hat{\mathbf{H}}^{\dagger} \boldsymbol{\Omega}^{-1} \hat{\mathbf{H}} \mathbf{P}^{1/2}\big)\big] \tag{2.167}$$

where the expectation is with respect to the channel estimates. Since both the received signal and the equivalent noise terms increase with user power, the capacity is upper bounded by

$$\overline{\mathcal{C}}_{\text{SUM}} < \mathbb{E}_{\mathbf{E}}\big[\log_2 \det\big(\mathbf{I}_K + \mathbf{P}^{1/2} \hat{\mathbf{H}}^{\dagger} \Big(\sum_{k=1}^{K} p_k \mathbf{C}_k\Big)^{-1} \hat{\mathbf{H}} \mathbf{P}^{1/2}\big)\big], \tag{2.168}$$

which is tight at high SNR ($\rho \to \infty$). At high transmit SNR capacity can therefore only be increased by increasing the SNR of the channel estimates – for example by increasing the SNR of the uplink pilots.

To aid future analysis it is useful to define an equivalent 'whitened' channel

$$\check{\mathbf{H}} = \boldsymbol{\Omega}^{-1/2} \hat{\mathbf{H}}, \tag{2.169}$$

so that (2.167) becomes simply

$$\overline{\mathcal{C}}_{\text{SUM}} = \mathbb{E}_{\mathbf{E}}\big[\log_2 \det\big(\mathbf{I}_K + \rho \mathbf{P}^{1/2} \check{\mathbf{H}}^{\dagger} \check{\mathbf{H}} \mathbf{P}^{1/2}\big)\big]. \tag{2.170}$$

This can be thought of as the result of applying a whitening transform to the received signal,

$$\check{\mathbf{y}} = \boldsymbol{\Omega}^{-1/2} \mathbf{y} \tag{2.171}$$

$$= \check{\mathbf{H}} \mathbf{P}^{1/2} \mathbf{x} + \check{\boldsymbol{\nu}} \tag{2.172}$$

where $\mathbb{E}\big[\check{\boldsymbol{\nu}}\check{\boldsymbol{\nu}}^{\dagger}\big] = \mathbf{I}_M$, a weighting that favours the signal 'directions' that (on average) contain

lower amount of channel estimation error. Whitened channel vectors may also be defined

$$\check{\mathbf{h}}_k = \mathbf{\Omega}^{-1/2}\mathbf{h}_k. \tag{2.173}$$

Under linear detection the estimate of user symbol k is,

$$\hat{x}_k = \mathbf{w}_k^T\mathbf{y} \tag{2.174}$$

$$= \sqrt{p_k}\mathbf{w}_k^T\hat{\mathbf{h}}_k x_k + \mathbf{w}_k^T\Big(\sum_{j\neq k}\sqrt{p_j}x\hat{\mathbf{h}}_j x_j + \sum_{j=1}^{K}\sqrt{p_j}\mathbf{e}_j x_j + \boldsymbol{\eta}\Big). \tag{2.175}$$

By the same reasoning as above, the expected user capacity can be written [15]

$$\overline{\mathcal{C}}_k = \mathbb{E}_{\mathbf{E}}\big[\log_2\left(1+\mathrm{SINR}_k^{\mathrm{UL}}\right)\big] \tag{2.176}$$

with effective SINR

$$\mathrm{SINR}_k^{\mathrm{UL}} = \frac{\rho_k|\mathbf{w}_k^T\hat{\mathbf{h}}_k|^2}{\sum_{j\neq k}\rho_j|\mathbf{w}_k^T\hat{\mathbf{h}}_j|^2 + \mathbf{w}_k^T\mathbf{\Omega}\mathbf{w}_k^*}. \tag{2.177}$$

The estimator that minimises mean square error with respect to noise and random channel estimation error is given by [1]

$$\mathbf{W}^{(\mathrm{MMSE})} = \mathbf{P}^{-1/2}\Big(\check{\mathbf{H}}^\dagger\check{\mathbf{H}} + \rho^{-1}\mathbf{P}^{-1}\Big)^{-1}\check{\mathbf{H}}^\dagger\mathbf{\Omega}^{-1/2}. \tag{2.178}$$

Figure 2.18 shows the mean sum capacity bound under both optimal and linear detection with varying CSI quality (defined by the pilot strengths, ρ_{CSI}), in the i.i.d Rayleigh channel.

Downlink Channel Capacity

On the downlink, the received signal at user k is

$$y_k = \hat{\mathbf{h}}_k^T\mathbf{x} + \mathbf{e}_k^T\mathbf{x} + \eta. \tag{2.179}$$

Assuming linear precoding is used,

$$y_k = \sum_{j=1}^{K}\hat{\mathbf{h}}_k^T\mathbf{w}_j s_j + \theta_k + \eta, \tag{2.180}$$

where θ_k is the component through the unknown channel

$$\theta_k = \sum_{j=1}^{K}\mathbf{e}_k^T\mathbf{w}_j s_j, \tag{2.181}$$

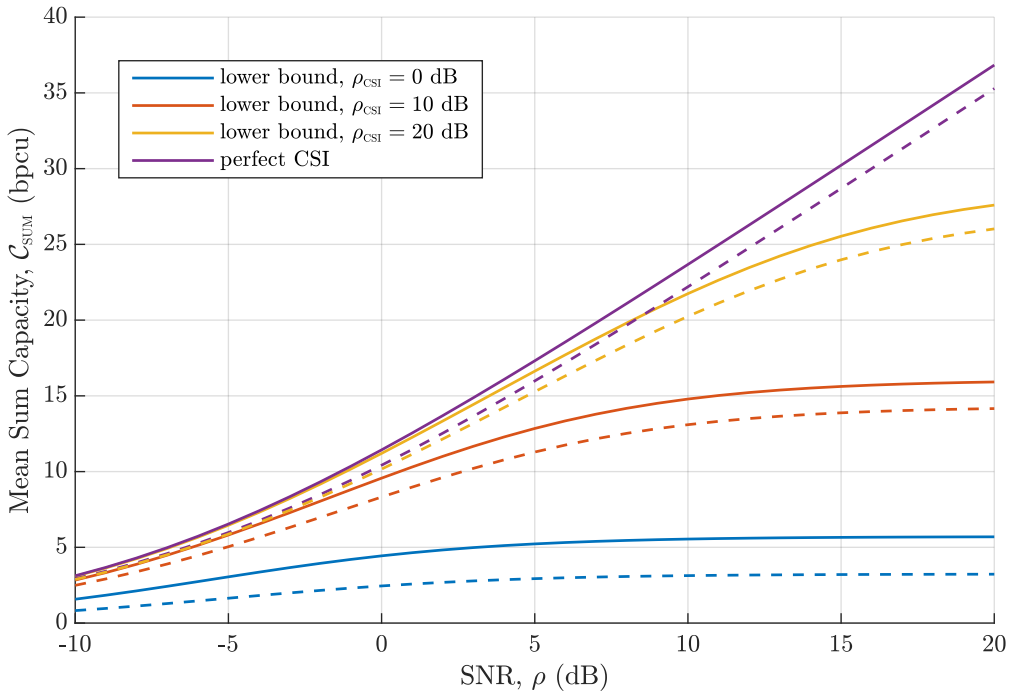

Figure 2.18: Uplink channel capacity with varying quality of CSI, $M = 8, K = 4$. Solid lines: optimal detection, dotted lines: MMSE detection.

with

$$\begin{aligned}\mathbb{E}_{\mathbf{e}_k}\left[\theta_k^* \theta_k\right] &= \sum_{j=1}^{K} \rho \mathbf{w}_j^\dagger \mathbb{E}\left[\mathbf{e}_k^* \mathbf{e}_k^T\right] \mathbf{w}_j \\ &= \sum_{j=1}^{K} \rho \mathbf{w}_j^T \mathbf{C}_k \mathbf{w}_j^*.\end{aligned} \tag{2.182}$$

This results in an effective downlink SINR

$$\text{SINR}_k^{\text{DL}} = \frac{\rho |\hat{\mathbf{h}}_k^T \mathbf{w}_k|^2}{\rho \sum_{j \neq k} |\hat{\mathbf{h}}_k^T \mathbf{w}_j|^2 + \rho \sum_{j=1}^{K} \mathbf{w}_j^T \mathbf{C}_k \mathbf{w}_j^* + 1}. \tag{2.183}$$

2.4.3 Power Control

The distribution of users within the propagation environment means that the proper application of power control is necessary to achieve good performance. In its simplest form, power control counteracts the pathloss between users and BS to ensure a signal of appropriate strength is received for each user. More advanced power control schemes may allocate user power levels to meet specific quality of service targets, noting the dependencies of sum & user capacities on power control coefficients above.

A large variety of cellular power control algorithms have been studied, operating with differing targets and constraints [37]. Here two power control schemes employed in the research are given – average received power control and max-min power control.

Average Received Power

The simplest form of cellular power control acts to mitigate the varying pathloss between the BS and user to achieve a target average received power level. On the uplink, for example, when the channels from user k to each BS antenna have the same large scale fading,

$$p_k \mathbb{E}\left[\|\mathbf{h}_k\|^2\right] = p_k \beta_k M. \tag{2.184}$$

Since the power control coefficients depend only on large scaling fading (pathloss), they are invariant to small scale fading across time and frequency, and can be kept constant across multiple coherence blocks, and therefore need not be frequently updated.

In much of the literature power control is performed so that each channel has an average gain of 1,

$$p_k = \frac{1}{\beta_k}. \tag{2.185}$$

The SNR parameter ρ used as a variable parameter when analysis how performance varies with transmit power. This method is employed for uplink transmission in the work in this thesis.

Max-Min Power Control

More sophisticated methods modify the power control based on instantaneous CSI towards a specific performance target. Such methods may be used for downlink transmission, where CSI is available at the BS. Max-min power control maximises the minimum of the K user SINRs, subject to a total transmit power constraint. It is straightforward to show that under max-min power control, the SINR of all users is equal [135], and thus this power control method ensures a 'fair' distribution of power between users.

The downlink max-min power control optimisation is

$$\begin{aligned} \underset{\rho_j \forall j}{\text{maximise}} \quad & \min_k \text{SINR}_k \\ \text{subject to} \quad & \sum_{k=1}^{K} \rho_k \leq P_T. \end{aligned} \tag{2.186}$$

Often the SINR can be written in the form

$$\text{SINR}_k = \frac{\rho_k}{a_k + \sum_{j=1}^{K} b_{k,j} \rho_j}, \tag{2.187}$$

e.g. (2.124) & (2.183), and the optimisation can be reformulated as a convex problem and efficiently solved. The convex problem is

$$\begin{aligned} \underset{\rho_j \forall j}{\text{minimise}} \quad & \psi \\ \text{subject to} \quad & t_k \leq \psi \\ & \sum_{j=1}^{K} \rho_j \leq P_T. \end{aligned} \tag{2.188}$$

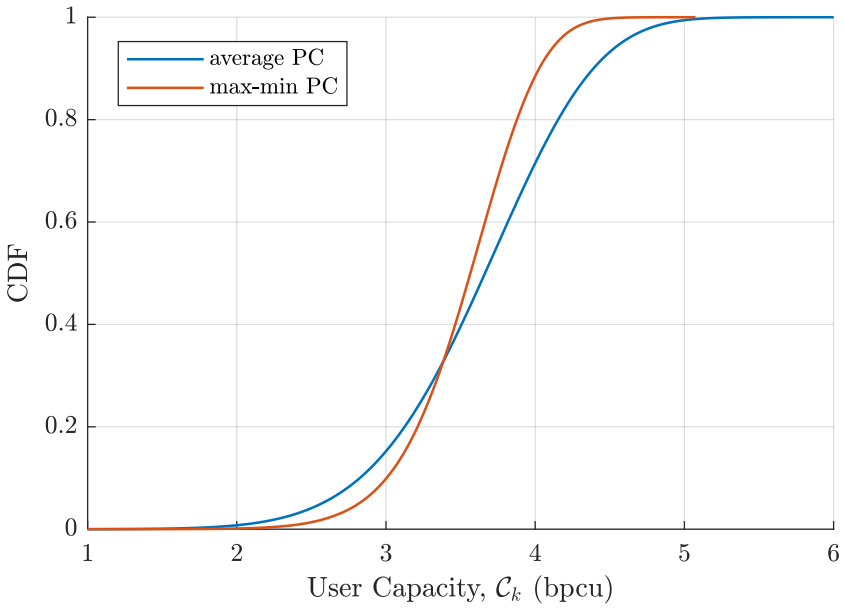

Figure 2.19: Distribution of downlink user capacities under average and max-min power control, ZF beamforming, $M = 8, K = 4, P_T = 10$ dB, i.i.d Rayleigh fading.

where ψ is a dummy variable and t_k a posynomial

$$\begin{aligned} t_k &= \frac{1}{\text{SINR}_k} \\ &= a_k \rho_k^{-1} + \sum_{j=1}^{K} b_{k,j} \rho_j \rho_k^{-1}. \end{aligned} \tag{2.189}$$

This can then be solved using geometric programming [135]. Uplink-downlink duality [218] can then be exploited to update the beamforming vectors under the new power allocations, repeating iteratively to find the optimal pairings.

When ZF beamforming is used – which is optimal at high transmit power – the transmit signal under max-min power control is simply given by,

$$\mathbf{x} = \gamma \mathbf{W}^{(\text{ZF})} \mathbf{s}, \tag{2.190}$$

where

$$\gamma = \frac{1}{\left\| \mathbf{W}^{(\text{ZF})} \right\|} \sqrt{\frac{P_T}{\rho}}. \tag{2.191}$$

Figure 2.19 illustrates the difference between the two power control schemes, showing the distribution of downlink user capacities under ZF beamforming in the i.i.d Rayleigh fading channel with $M = 8, K = 4$. With average power control, power is allocated to each user stream based only on large scale fading (which here is equal for all users), whereas with max-min power control, the power is allocated such that each user experiences the same capacity. Under max-min power control a tighter range of capacities are seen – the worst user capacities are improved at the expense of the best user capacities. A related power control scheme finds

the power allocations required to achieve target user SINRs [176], but is outside the scope of this work.

2.5 Advanced MU-MIMO Architectures

Whilst the potential benefits of MU-MIMO technology have been well known since the 1990s, they have yet to be properly realised in cellular systems. Despite its inclusion in fourth (LTE) generation cellular standards, the technique was not widely implemented due to limitations imposed by the standards, stemming from some key practical challenges:

- the need for full downlink CSI at the BS for downlink precoding. European & North American LTE implementations operate using frequency divison duplex (FDD), where downlink and uplink transmission occur in different frequency bands, and downlink CSI must therefore be estimated using downlink pilots and fed back to the BS before transmission occurs (cf. Section 2.4.2). Early LTE standard releases opted to instead use rudimentary codebook-based precoding, where the precoding matrices are selected from a set of fixed candidates (rather than calculated from CSI), at the cost of poor performance [122]. Later LTE-Advanced releases exploit full downlink CSI, but the overheads associated with feeding back CSI limit the number of BS antennas that can be deployed and the number of users served [117].

- the prohibitive complexity of optimal non-linear MIMO processing, and poor performance of lower complexity linear processing when the channel is poorly conditioned. In conjunction with the limits imposed by CSI, this means that only modest sum spectral efficiencies are achieved in LTE-A [22].

- the poor performance of users at the cell-edge. Cellular BS operate with transmit power constraints, and MU-MIMO operation requires this power to be split between multiple users, which can lead to poor SINR at the cell-edge, where pathloss is greater and interference from neighbouring cells higher [122].

As spectrum becomes increasingly scarce, fifth generation and future wireless systems must employ MU-MIMO technology much more effectively if they are to achieve the high spectral efficiencies required to meet growing traffic demands. The last 10 years has seen significant advances towards practical forms of effective MU-MIMO technology, with much of it focused around two general architectures: massive MIMO and distributed MIMO.

2.5.1 Massive MIMO

A 'massive' MU-MIMO system is one in which a large number of antennas are deployed at the BS (typically in an array), greatly exceeding the number of active users being served, $M \gg K$. Whilst strictly just a specific configuration of conventional MU-MIMO, massive – or 'large-scale' – MIMO has attracted great interest within both the research community and industry due to a number of practical performance benefits that emerge from the properties of the MIMO propagation channel as the number of antennas asymptotically approaches infinity

[133]. Fundamental to the original massive MIMO concept is TDD operation, such that the uplink and downlink channels are reciprocal and downlink channel estimates can be obtained from uplink pilots (cf. Section 2.4.2) – enabling the number of BS antennas to be increased without increasing the signalling overheads associated with obtaining full CSI.

The benefits provided by massive MIMO give it the potential to overcome the limitations of previous cellular MU-MIMO implementations to achieve order-of-magnitude improvements in both spectral and energy efficiency [113]. It is now considered a core enabling technology for fifth generation cellular system, with elements of it already included in fifth generation standards [66].

Asymptotic Results

The core massive MIMO concept stems from two properties of the MIMO channel that are exhibited – under the right propagation conditions – as the number of BS antennas increases [133]:

- *channel hardening* – the strength of each user channel vector tends towards a deterministic constant that depends only on the channel pathloss.
- *asymptotic orthogonality* – the inner product between any two user channel vectors becomes vanishingly small to the user channel strengths.

The channel hardening follows directly from the law of large numbers,

$$\lim_{M\to\infty} \frac{1}{M}\|\mathbf{h}_k\|^2 = \frac{1}{M}\mathbb{E}\big[\|\mathbf{h}_k\|^2\big] = \beta_k. \tag{2.192}$$

The asymptotic orthogonality property relies on the propagation environment providing *favourable propagation* [135], such that the inner product between two channel vectors grows at a rate less than M,

$$\lim_{M\to\infty} \frac{1}{M}\mathbf{h}_k^\dagger \mathbf{h}_j = 0, \tag{2.193}$$

or, alternatively [15], that

$$\lim_{M\to\infty} \frac{\mathbf{h}_k^\dagger \mathbf{h}_j}{\|\mathbf{h}_k\|^2} = 0, \tag{2.194}$$

When the entries of the two user channel vectors are randomly distributed and uncorrelated this follows from the law of large numbers, but it also holds for non-random channels, e.g. line-of-sight [237]. For a fixed number of users K, the channel hardening and asymptotic orthogonality properties on the uplink are jointly captured by the limit

$$\lim_{M\to\infty} \frac{1}{M}\mathbf{H}^\dagger\mathbf{H} = \begin{bmatrix} p_1\beta_1 & & \\ & \ddots & \\ & & p_K\beta_K \end{bmatrix}, \tag{2.195}$$

visualised in Figure 2.20 for an i.i.d Rayleigh fading channel.

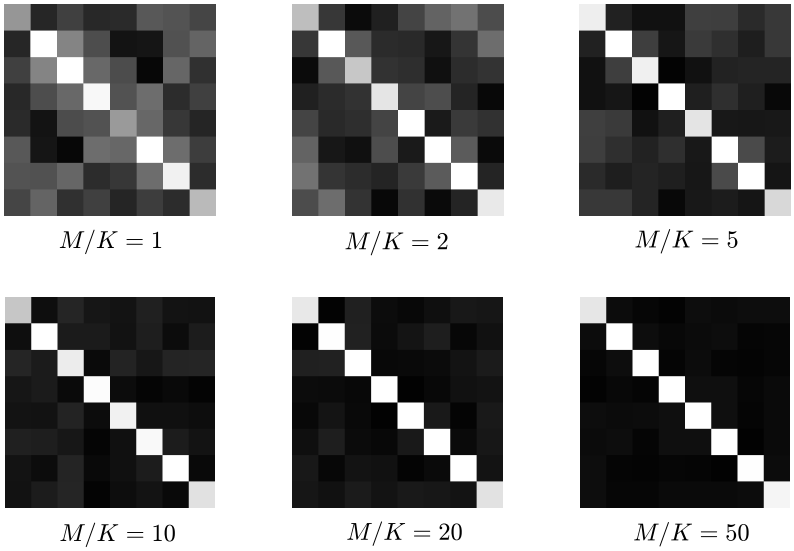

Figure 2.20: Intensity plots of $\frac{1}{M}\mathbf{H}^{\dagger}\mathbf{H}$ for various values of M, i.i.d Rayleigh fading channel, $K = 8$.

Benefits of Large Arrays

Clearly, no practical system can operate with an infinite number of antennas, but the asymptotic results are useful in giving insights into the behaviour of MIMO systems operating with a large array of antennas.

- **Linear processing becomes optimal**. As shown in Figure 2.15, under i.i.d Rayleigh fading increasing the number of BS antennas improves the MIMO channel condition number and performance of linear processing. More generally, when the user channels exhibit asymptotic orthogonality, inter-user interference can be eliminated by simple matched filter, zero-forcing, or MMSE detection. With SNR $\rho = \gamma/M$, by (2.195) the asymptotic sum capacity (under optimal MMSE-SIC detection) is given by

$$\lim_{M\to\infty} \log_2 \det\left(\mathbf{I}_K + \frac{\gamma}{M}\mathbf{H}^{\dagger}\mathbf{H}\right) = \sum_{k=1}^{K} \log_2\left(1 + \gamma p_k \beta_k\right), \tag{2.196}$$

  i.e. K parallel channels with $\mathrm{SINR}_k = \gamma p_k \beta_k$. These SINRs are also asymptotically achieved by MF detection,

$$\mathrm{SINR}_k^{(\mathrm{MF})} = \frac{\frac{\gamma}{M} p_k \|\mathbf{h}_k\|^2}{\frac{\gamma}{M}\sum_{j\neq k} p_j \frac{|\mathbf{h}_k^{\dagger}\mathbf{h}_j|^2}{\|\mathbf{h}_k\|^2} + 1} \to \gamma p_k \beta_k, \tag{2.197}$$

  which follows from (2.192)-(2.194), as well as under ZF & MMSE detection – which can be shown by substitution of (2.195) into (2.106) & (2.113). By uplink-downlink duality, this asymptotic optimality of linear processing also holds for downlink precoding.

  With a finite number of BS antennas, the user channels are generally not perfectly or-

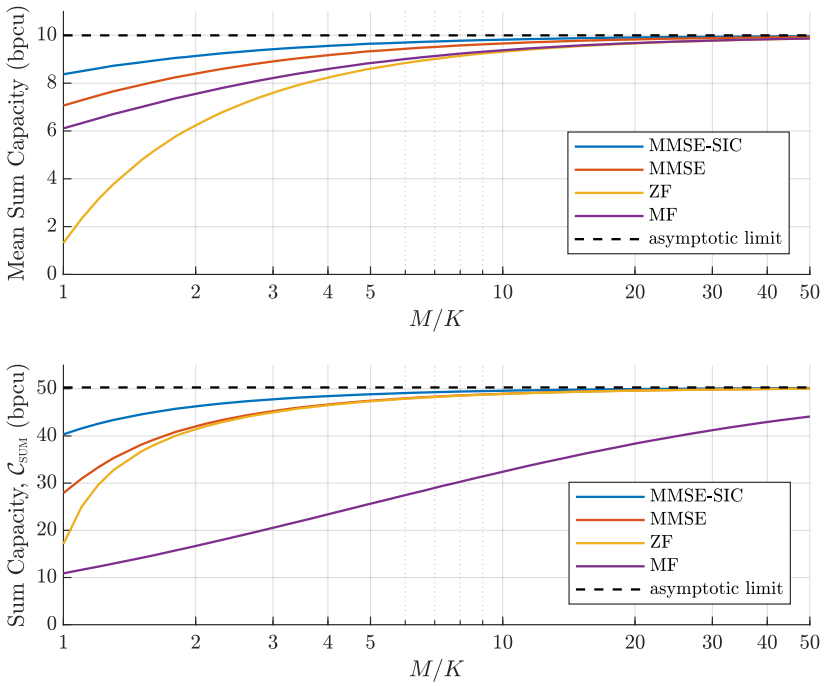

Figure 2.21: Mean sum capacity under linear detection with varying M, i.i.d Rayleigh fading, $K = 10$, $p_k\beta_k = 1$. Top: $\gamma = -5$ dB, bottom: $\gamma = 15$ dB.

thogonal, and linear processing does not fully achieve the sum capacity. However, as the channels decorrelate ($\mathbf{h}_k^\dagger \mathbf{h}_j / \|\mathbf{h}_k\|^2$ shrinks), the performance of linear methods improves. As per-the discussion in Section 2.3, at high SNR, MMSE and ZF perform well whilst MF produces high inter-user interference, whereas at low SNR MMSE and MF perform well, as shown in Figure 2.21.

- **The effects of fast fading disappear**. The SINRs that are asymptotically achieved, $\text{SINR}_k = \gamma p_k \beta_k$, are independent of the specific channel realisation, depending only on the channel pathloss and user power control. As a result the user capacities stay constant across multiple coherence blocks, despite the channel realisations changing to fast fading. With finite arrays, variations in capacity will tend to decrease as the number of BS antennas is increased, as shown in Figure 2.22 for the i.i.d Rayleigh fading channel. This can be interpreted from a diversity perspective – the diversity gain of $d = M - K + 1$ provided by MMSE detection 'averaging' over the effects of fast fading.

  As well as improving user outage capacities, the removal of fading effects has a wide range of potential benefits, such as simplifying resource allocation [113], enabling more advanced uplink power control methods [236] and removing the need for downlink pilots [235].

- **Transmit power reduces & energy efficiency increases**. The array gain provided by the antennas means that on the uplink the user transmit power can be reduced as $\rho = \gamma/M$ whilst maintaining the same average level of received power [212]

$$\frac{\gamma}{M} p_k \mathbb{E}\left[\|\mathbf{h}_k\|^2\right] = \gamma p_k \beta_k. \tag{2.198}$$

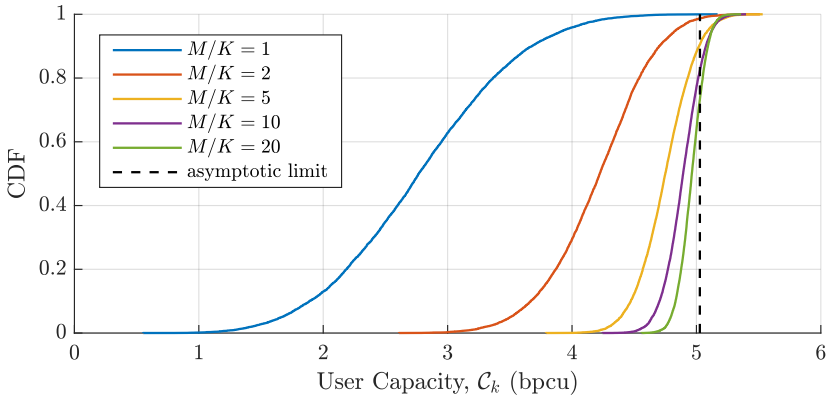

Figure 2.22: CDF of user capacities under MMSE detection with varying M, i.i.d Rayleigh fading, $\gamma = 10$ dB, $p_k\beta_k = 1$, $K = 10$.

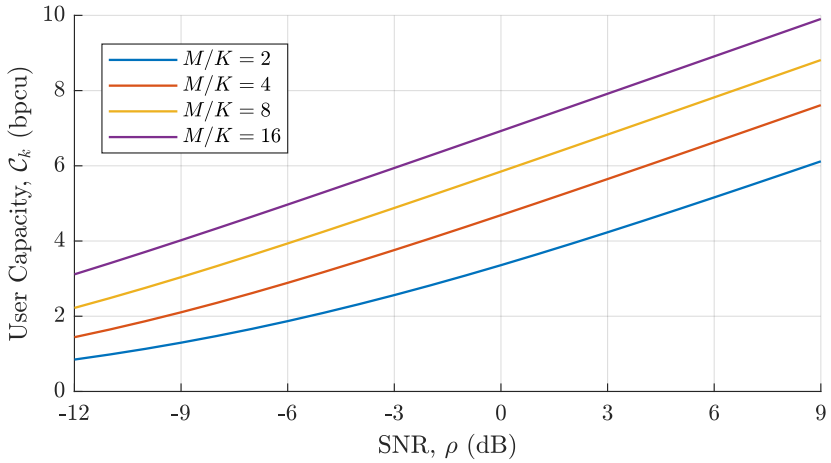

Figure 2.23: User capacity scaling with SNR for different numbers of antennas, i.i.d Rayleigh fading, $K = 8$, $p_k\beta_k = 1$.

Doubling the number of BS antennas enables the transmit power to be reduced by ~ 3 dB, whilst maintaining a similar level of performance[6], as shown in Figure 2.23.

The array gain is provided regardless of whether the propagation environment provides favourable propagation conditions, and is also achieved on the downlink, where the pre-coded signals constructively superimpose at the user, boosting received SNR. Along with improving the radiated energy efficiency, this means that the individual power requirement of each transmit RF chain is lowered, enabling the use of cheaper components [113].

Practical Massive MIMO Propagation

Since many of the benefits of massive MIMO are contingent on the presence of favourable propagation conditions, there has been considerable research into the performance achievable by massive MIMO systems operating in practical propagation environments with finite numbers of antennas.

[6]Under MMSE or ZF detection, the array gain provided is actually $M - K + 1$ [19].

Whilst much analysis assumes i.i.d Rayleigh fading, it has been shown that both LoS [155] and Rician channels [244] are also capable of providing favourable propagation, as long as the angular separation of the users is sufficient. Under the correlated fading model outlined in Section 2.4.1, it has been shown that spatially correlated fading can actually be beneficial, providing the eigenspaces of the channel correlation matrices do not significantly overlap [15]. There are well known benefits to scheduling users according to their channel correlations in conventional MU-MIMO systems [207], and this has also been shown to be beneficial under line-of-sight conditions in massive MIMO systems [237].

Measurements taken with uniform circular array in an urban environment with 2 users found that the measured user channel correlations were significantly higher than under i.i.d fading, but also decreased as the number of active antennas was increased – up to around $M = 20$ where they plateaued [60]. A second study using a 128 element uniform linear array and randomly selecting groups of 3 users (from 36 captured user channels) found that the correlation decreased significantly when increasing M from 20 to 128, for both LoS and non-LoS locations [166]. Using a 'virtual' ULA in an outdoor environment it was found in [93] that channel correlations and eigenvalue spread decreased with M, but that the gain from adding antennas decreased, with significant channel correlations still present with 112 antennas. Another measurement campaign found eigenvalue spreads and channel capacities for co-located users with non-LoS channels and separated users with LoS channels were consistent with i.i.d Rayleigh fading, but that for co-located users with non-LoS channels they were not [61].

The practical benefits of massive MIMO have been demonstrated by a number of real-time testbeds [130], [217], [193], [79]. Notably, 256QAM modulation was simultaneously transmitted by 22 users and accurately detected by a BS with 128 antennas in [79], and an uplink sum spectral efficiency of 79 bps/Hz was achieved by 12 users in [82], both in indoor rich scattering LoS environments. In [80], 8 users were simultaneously served with QPSK modulation on both uplink and downlink by a BS with 100 antennas, in an outdoor mobile environment. In these cases, channel estimation was performed by the users transmitting uplink pilots on different subcarriers within the same coherence block.

Research Challenges & Opportunities

The early work demonstrating the potential benefits of massive MIMO was followed by an explosion in research exploring the new opportunities and addressing the practical challenges associated with implementing it commercially.

Much work has focused on more accurately quantifying the performance of massive MIMO systems, for example for more realistic scenarios where the number of BS antennas is finite [154] and fading is correlated [94]. This research frequently departs from the approaches used previously for MU-MIMO analysis by making use of random matrix theory and asymptotic results that become reasonable approximations in the 'massive' regime. Beyond conventional single-cell analysis, the impact of inter-cell interference in non-cooperative multi-cell configurations has also been considered, for example [18].

The overheads associated with the uplink pilots used for channel estimation were identified early on as a potential limiting factor – the resources required growing with the number of users

being served. The seminal massive MIMO work identified a 'pilot contamination' problem, in which the re-use of piloting resources in adjacent cells led to contaminated channel estimates that meant inter-cell interference (from simultaneous transmissions in the adjacent cells) remained after inter-user interference within the cell had been cancelled [133]. This led to a glut of research on pilot design & allocation and mitigation techniques [54]. More recently it has been shown that pilot contamination does not fundamentally limit performance if the second order user channel statistics of the contaminating users are known during MMSE channel estimation and signal detection/precoding [103], providing some mild conditions on the channel covariances are met [17].

Initially, there was scepticism about the feasibility of implementing massive MIMO, due to perceived issues with cost & complexity [19]. Numerous hardware implementations have since demonstrated that linear MIMO detection methods for relatively large systems are well within the capabilities of current digital signal processing platforms [229], [229]. The conventional massive MIMO architecture also relies on each antenna being connected to its own individual RF chain, with digital to analogue convertor (DAC) and power amplifier (PA) for the downlink and low noise amplifier (LNA) and analogue to digital convertor (ADC) for the uplink, leading to concerns about hardware cost & power consumption as the array size grows. As a result, there has been significant investigation into the use of hybrid architectures, where analogue beamforming at the antennas is employed to reduce the number of RF chains required. However, effective use of hybrid methods in wideband systems at sub-6GHz is challenging due to the reduced flexibility of analogue beamforming compared to digital per-subcarrier processing [143].

An alternative – and also widely investigated – approach to reducing the complexity of massive MIMO hardware is to use lower precision hardware in the transceivers, to reduce cost and/or improve energy efficiency. This is based on the observation that often non-linearities introduced due to hardware impairments will 'average' across antennas, and their effects diminish as the number of antennas is increased [16], and has led to particular interest in the use of reduced resolution ADCs and DACs. From an academic perspective, the study of 1-bit ADCs poses interesting technical challenges [116], [146], but a more thorough consideration of energy efficiency indicates that the use of 4-8 bit resolution ADCs is preferable [184].

Beyond the topics briefly outlined here, a huge amount of research has considered other novel applications and analyses of massive MIMO technology. One interesting area is the exploitation of the *excess degrees of freedom* provided by deploying a large excess of BS antennas [113]. Chapter 3 the use of these degrees of freedom to reduce the peak-to-average power ratio (PAPR) of the massive MIMO transmit signals and further relax hardware requirements and improve system energy efficiency.

2.5.2 Distributed MIMO C-RAN

A 'distributed' MU-MIMO system has BS antennas that are distributed geographically within the cell or service area, rather than being co-located in an array at a single site as in a conventional MU-MIMO system. This architecture has long been of interest to researchers, and can be seen as the convergence of two concepts:

1. The co-operation of different BS to reduce or eliminate inter-cell interference and improve

the performance of cell-edge users. This idea predates third generation systems [176], and is based on two or more BS jointly transmitting and/or detecting user data across cell boundaries, using beamforming & shared CSI, such that inter-cell interference contributes constructively rather than destructively [114]. This is particularly attractive as cellular deployments densify with the use of small cells, since the reduced pathloss means that interference otherwise becomes the capacity-limiting factor [63]. A basic level of BS co-operation is included in LTE-A as co-ordinated multipoint (CoMP) [104].

2. The distribution of BS antennas in a MU-MIMO or massive MIMO system to improve geographical coverage. Distributed antenna systems were originally conceived of as a means for improving indoor cellular coverage by providing *macro-diversity* gain to counteract the effects of high pathloss, for example due to blockage [181]. The use of distributed antennas in conjunction with MU-MIMO has since been considered [195], and shown to provide benefits in terms of both mean and outage capacities compared to centralised configurations [86].

For fifth generation cellular systems there has been significant interest in the use of a so-called cloud (or centralised) radio access network (C-RAN) architecture, in which the signal processing and network functions for multiple base stations – or 'remote radio heads' – is performed at a single central processor (CP) [168], connected via fronthaul connections. The C-RAN architecture is potentially a key technology for unlocking the full benefits to distributed MIMO, enabling the group of remote radio heads to be treated as a single 'network' MIMO system. Uplink detection and downlink precoding can then be performed jointly across all distributed antennas, allowing full control of the inter-user interference whilst capturing all diversity and macro-diversity present in the network [63].

At its most extreme, distributed MIMO could be used to eliminate traditional cell boundaries, with a large number of radio heads jointly serving all active users across an extended macro-coverage area. This 'cell-free' concept is not new [195], but has recently been extended to incorporate ideas from massive MIMO, so that channel hardening and favourable propagation can be exploited alongside macro-diversity in order to deliver a uniform quality of service to all active users [210], [153]. The early work on cell-free massive MIMO focused on a fully distributed architecture (many single-antenna radio heads), and the use of simple decentralised detection and precoding based on matched filtering [152]. However, it has since been shown that using a reduced number of multi-antenna radio heads is a more practical and effective way of capturing channel hardening [35], and that significant performance benefits result from using more complex centralised processing [20].

Despite recent progress, there are a number of practical challenges that remain to be solved before the full potential of distributed MIMO can be realised.

System Model & Simulation Configurations

This thesis considers distributed MIMO C-RAN systems in which L remote receivers/transmitters each equipped with M antennas jointly serve K active users, as shown in Figure 2.24.

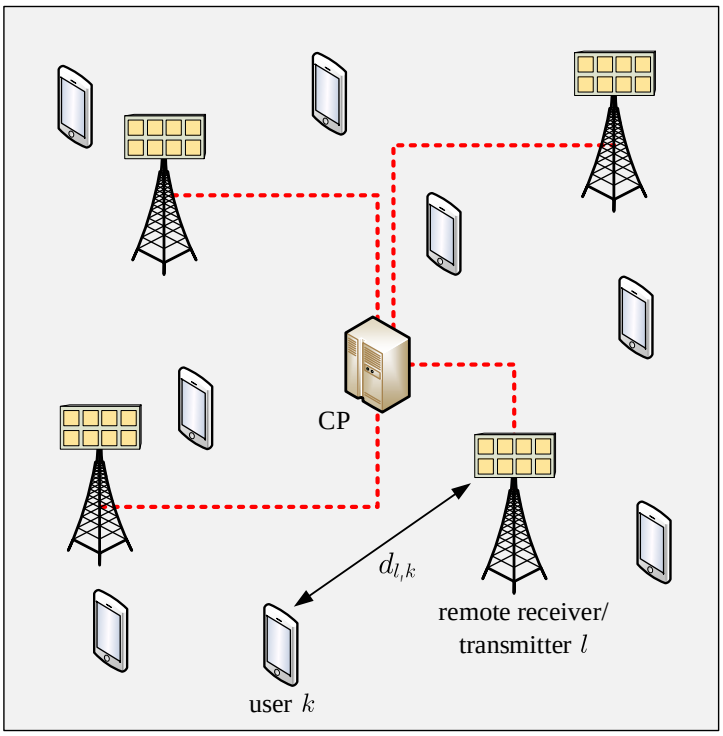

Figure 2.24: Example distributed MIMO C-RAN configuration with $K = 8, L = 4, M = 8$.

On the uplink, the received signal at receiver l is given by the standard MIMO uplink relation,

$$\mathbf{y}_l^{\mathrm{UL}} = \mathbf{H}_l^{\mathrm{UL}} \mathbf{x}^{\mathrm{UL}} + \boldsymbol{\eta} \tag{2.199}$$

where $\mathbf{x} \sim \mathcal{CN}\big(0, \rho \mathbf{I}_K\big)$, $\boldsymbol{\eta} \sim \mathcal{CN}\big(0, \mathbf{I}_M\big)$, and

$$\mathbf{H}_l^{\mathrm{UL}} = \Big[\sqrt{p_1}\mathbf{h}_{l,1} \ldots \sqrt{p_K}\mathbf{h}_{l,K}\Big] \tag{2.200}$$

with $\mathbf{h}_{l,k}$ the propagation channel between user k and receiver l, and p_k the user power control coefficient. These user channels can be modelled by appropriate fading and pathloss models, in the same way as conventional MU-MIMO channels. Note that here it is assumed that each of the K users has a channel to each of the L receivers; in reality some of these channels may be very weak if there is significant distance or blockage between a user and receiver, and hence some columns of $\mathbf{H}_l$ may have entries close to zero.

For numerical studies, a dense urban deployment is considered, in which the remote receivers/transmitter and users are randomly distributed within a 200 m $\times$ 200 m area. All users are at a height of 1 m, and remote antennas at a height of 6m. Pathloss is modelled as described in Section 2.1.4, using pathloss exponent $\gamma = 2.9$, $\sigma_\Psi^2 = 5.7$ dB log-normal shadow fading, and a 3.5 GHz carrier frequency [204]. User uplink power control, p_k, is applied such that the average received power for each user is the same

$$\frac{1}{ML}\mathbb{E}\Big[\sum_{l=1}^{L} p_k \big\|\mathbf{h}_{l,k}\big\|^2\Big] = \frac{1}{L}\sum_{l=1}^{L} p_k \beta_{l,k} = 1, \tag{2.201}$$

in order to normalise user channels for comparison purposes whilst preserving the differences in large scale fading that characterise the distributed MIMO propagation channels [175]. Mean capacities are calculated by averaging over both user/receiver/transmitter positions and channel fading realisations.

On the downlink, the received signal at user k is

$$y_k^{\mathrm{DL}} = \sum_{l=1}^{L} \mathbf{h}_{l,k}^{\mathrm{DL}} \mathbf{x}_l^{\mathrm{DL}} + \eta \tag{2.202}$$

where $\mathbf{x}_l^{\mathrm{DL}}$ is the precoded signal transmitted by transmitter l. This can also be written in matrix form

$$\mathbf{y}^{\mathrm{DL}} = \sum_{l=1}^{L} \mathbf{H}_l^{\mathrm{DL}} \mathbf{x}_l^{\mathrm{DL}} + \boldsymbol{\eta}. \tag{2.203}$$

Distributed MIMO Uplink

On the uplink, the ensemble of all received signals can be represented a global or network level by a single uplink MIMO equation

$$\mathbf{y}_{\mathrm{G}} = \mathbf{H}_{\mathrm{G}} \mathbf{x} + \boldsymbol{\eta} \tag{2.204}$$

where

$$\mathbf{y}_{\mathrm{G}} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_L \end{bmatrix}, \qquad \mathbf{H}_{\mathrm{G}} = \begin{bmatrix} \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_L \end{bmatrix}. \tag{2.205}$$

The sum MIMO uplink capacity is given by

$$\mathcal{C}_{\mathrm{SUM}}^{\mathrm{UL}} = \log_2 \det \left(\mathbf{I}_K + \rho \mathbf{H}_{\mathrm{G}}^{\dagger} \mathbf{H}_{\mathrm{G}} \right) \tag{2.206}$$

$$= \log_2 \det \Big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \mathbf{H}_l^{\dagger} \mathbf{H}_l \Big), \tag{2.207}$$

and is achieved by *jointly* detecting the user symbols using all the received signals and MMSE-SIC detection. Joint linear MMSE detection can also be used,

$$\hat{\mathbf{x}} = \sum_{l=1}^{L} \mathbf{W}_l \mathbf{y}_l, \tag{2.208}$$

with detection matrices

$$\mathbf{W}_l = \rho \Big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \mathbf{H}_l^{\dagger} \mathbf{H}_l \Big)^{-1} \mathbf{H}_l^{\dagger}. \tag{2.209}$$

Observe that calculating the detection matrices for joint MMSE or MMSE-SIC symbol detection requires that the CP have access to *global* CSI. Various other detection methods have also been proposed that used distributed processing to reduce the CSI requirements, e.g. [11], [20], but come at the cost of reduced performance.

Figure 2.25 demonstrates how distributing the antennas leads to an improvement in both mean and outage capacity, for a configuration with $K = 4$ users, a total of $ML = 16$ antennas,

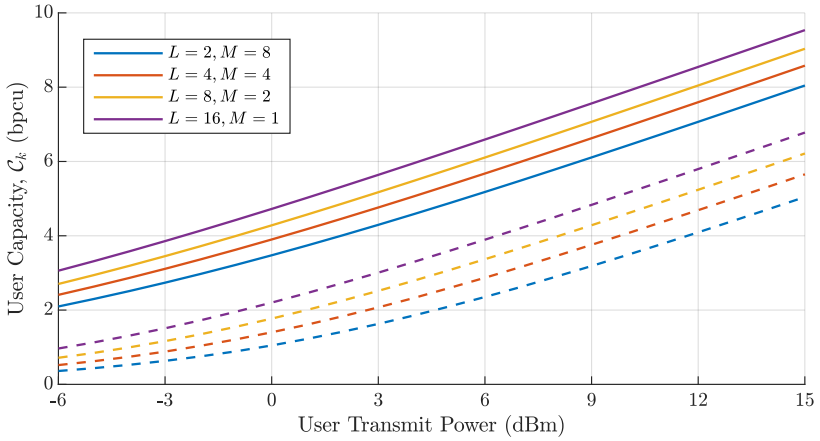

Figure 2.25: User uplink mean capacities for varying number of distributed receivers, L, i.i.d Rayleigh fading, $K = 4$, $ML = 16$. Solid line: mean capacity, dashed line: 10% outage capacity.

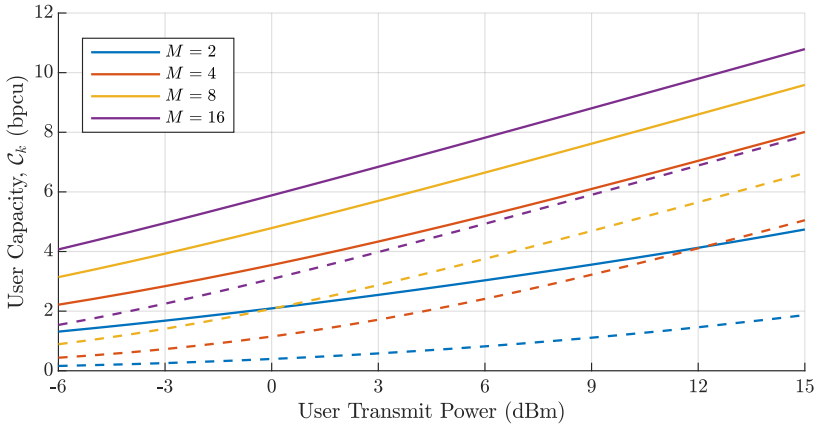

Figure 2.26: User uplink capacities for varying number of antennas per receiver, i.i.d Rayleigh fading, $K = 8$, $L = 4$. Solid line: mean capacity, dashed line: 10% outage capacity.

and MMSE detection. In practice, whilst using a large number of individual distributed antennas may be feasible in certain scenarios, e.g. when they can be spaced along a structure [59], the cost associated with deploying and connecting a large number of distinct remote radio heads at different sites may be prohibitive, and the use of a smaller number of multi-antenna radio heads might be an attractive compromise. This use of multiple antennas at each receiver provides both array gain and diversity, boosting mean and outage capacities as shown in Figure 2.26 Analysis can be readily extended for the case of imperfect CSI using the method outlined in Section 2.4.2.

Distributed MIMO Downlink

Whilst under joint processing the distributed MIMO uplink can be treated as a conventional MU-MIMO system, on the distributed MIMO downlink it is necessary to replace the total transmit power constraints used in conventional systems with per-transmitter power constraints.

This use of per-transmitter power constraints means that the duality used to establish the sum capacity and transmission strategies in Section 2.3 no longer applies, complicating the design and analysis of optimal precoding schemes.

As with conventional systems, the use of linear precoding is desirable from an implementation perspective,

$$\mathbf{x}_l = \mathbf{W}_l \mathbf{P}^{1/2} \mathbf{s}, \tag{2.210}$$

with precoding matrices, $\mathbf{W}_l$, and user power control, $\mathbf{P}$, chosen to ensure the per-transmitter power constraints are met

$$\mathbb{E}\big[\|\mathbf{x}_l\|^2\big] = \rho \operatorname{Tr}\big(\mathbf{W}_l \mathbf{P} \mathbf{W}_l\big) \leq P_T. \tag{2.211}$$

Under joint signal processing, ZF precoding with the conventional Moore-Penrose pseudo-inverse is a straightforward and popular strategy [96],

$$\mathbf{W}_l = \mathbf{H}_l^{\dagger} \big(\sum_{i=1}^{L} \mathbf{H}_i \mathbf{H}_i^{\dagger} \big)^{-1}. \tag{2.212}$$

Under max-min power control, the power must be backed off at each receiver in line with the precoding matrix with the highest power,

$$\mathbf{x}_l = \frac{\sqrt{P_T/\rho}}{\max\limits_j \|\mathbf{W}_j\|} \mathbf{W}_l \mathbf{s}, \tag{2.213}$$

meaning that only one of the transmitters will operate at full transmit power, with others potentially transmitting well below their maximum level. This contrasts with the case of a single total power constraint, where precoding with the Moore-Penrose pseudo-inverse uses all of the available power [220].

When the total number of transmit antennas exceeds the number of users, $ML > K$, an infinite number of ZF precoding matrices that perfectly eliminate inter-user interference exist – some of which may outperform the Moore-Penrose under per-transmitter constraints. Unfortunately, finding the optimal ZF precoding matrix generally involves the use of numerical methods. Under max-min power control, the optimal ZF precoding matrix[7] can be found by adapting the convex optimisation in [108],

$$\underset{\varphi, \mathbf{W}_l}{\text{maximise}} \quad \varphi \tag{2.214}$$

$$\text{subject to} \quad \sum_{l=1}^{L} \mathbf{H}_l \mathbf{W}_l \geq \varphi \mathbf{I}_K, \tag{2.215}$$

$$\Im\Big(\sum_{l=1}^{L} \mathbf{H}_l \mathbf{W}_l \Big) = \mathbf{0}, \tag{2.216}$$

$$\big\|\mathbf{W}_l\big\|^2 \leq P_T \quad \forall l, \tag{2.217}$$

where φ represents the received signal strength at all users.

[7] Here the power control is absorbed into $\mathbf{W}_l$.

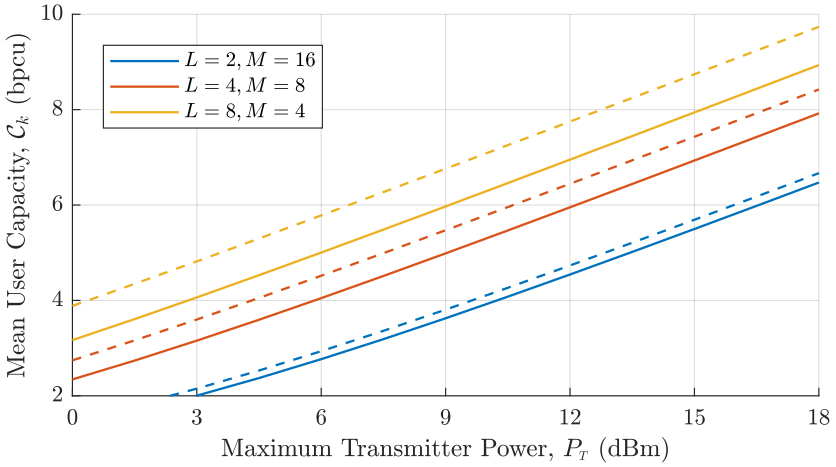

Figure 2.27: Mean downlink user capacity under ZF precoding with max-min power control, i.i.d Rayleigh fading, $K = 8$, $L = 4$. Solid line: standard ZF, dashed line: ZF-MPC.

Figure 2.27 compares the mean user capacities achieved under per-transmitter power constraints for the two ZF precoding schemes. In the first case the standard Moore-Penrose precoding matrices are used, with power control applied to ensure per-transmitter contraints are met. The second case (ZF-MPC) directly optimises the precoding matrices under multiple per-transmitter power constraints as in (2.214). The ZF-MPC precoding matrices better utilise the available power to increase the signal strength at the users, with performance improving relative to standard ZF as the number of transmitter power constraints increases. In practice, however, the standard ZF solution may be preferable due to the significantly higher computational complexity of ZF-MPC. There has been significant research into other downlink transmission schemes, and it has been shown that a variation of duality can also be exploited under per-transmitter power constraints to full characterise the achievable downlink capacities [242]. Approaches that use distributed processing rather than joint processing have also been proposed [153], [21].

Signalling & Co-ordination Challenges

Beyond the fundamental signal processing operations that must be adapted for distributed MIMO C-RAN, there are also key challenges related to the signalling & co-ordination required between the remote radio heads, CP and users:

- Joint detection and transmission requires the synchronisation of the clocks used in the receivers & transmitters, as well as the users. Whilst distributed synchronisation methods have been developed [197], the use of centralised synchronisation using either a shared control channel or the global navigation satellite system may be most appropriate in a C-RAN system, as discussed in [115] (and references therein).

- Joint processing requires full CSI on both the uplink and downlink. For the uplink, this can be obtained using pilots transmitted by the users, as described in Section 2.4.2. These estimates can either be formed locally at the receivers and transferred over fronthaul to

the CP, or the received pilot signals can be forwarded directly over fronthaul and the estimates formed at the CP. For the purposes of the work here, the former is assumed, as the availability of local CSI at each receiver enables some signal processing operations to be performed there. On the downlink, TDD operation is desired as this enables uplink pilots & reciprocity to be used [23], overcoming the bottleneck associated with CSI feedback in frequency division duplexing.

- Joint processing requires the transfer of both CSI and uplink/downlink data payloads over the fronthaul connections between CP and remote units. When these fronthaul connections are made by dedicated high capacity fibre this is straightforward, but in many scenarios the fronthaul may have limited capacity, and hence require the use of bespoke signal compression/quantization to maximise performance under this constraint [98].

The research in Chapters 4 & 5 of this thesis focuses specifically on final point, and the design of signal compression schemes that can enable high user data throughput to be achieved when the fronthaul capacity is limited.

2.6 Conclusion

Multi-user MIMO is a physical layer communication technology that provides significant spectral efficiency improvements by spatially multiplexing multiple users on the same time-frequency resource, and is expected to play a growing role in future wireless networks. This background chapter has provided a brief introduction into MU-MIMO systems, and acts as a reference for the mathematical models and ideas used in the MU-MIMO research in Chapters 3, 4 & 5.

The fundamental communication techniques that MU-MIMO systems are built on have been outlined – modulation, channel coding & OFDM – along with the principal signal processing operations: uplink multi-user signal detection, downlink signal precoding, power control & channel estimation. Important communication concepts that are used to guide the design & analysis of MU-MIMO systems have also been introduced – capacity, fading, spatial multiplexing, diversity, outage & array gain, with numerical examples provided throughout to illustrate key concepts and results. A variety of mathematical expressions that are re-used in the following research chapters have also been provided.

The final section of the chapter contains an overview of the massive MIMO and distributed MIMO architectures that are the focus of the research in this thesis. The massive MIMO section examines the benefits of deploying a large number of antennas at the MU-MIMO base station, before the distributed MIMO section discusses the key implications of geographically distributing base station antennas within the service area. This background material is supplemented by more in-depth reviews of the relevant state-of-art research provided at the beginning of each research chapter.

The background material provided in this chapter is not exhaustive, and largely limits it focus to those ideas and methods required to understand the subsequent research material. For the interested reader, [174], [212] & [15] represent excellent references for developing a further understanding of MU-MIMO systems.

Chapter 3

Clipping-based PAPR Reduction for the Massive MIMO Downlink

In fourth generation LTE systems it is estimated that BSs account for 80% of total network power consumption, of which between 30% and 60% can be attributed to the RF transmit frontend [7]. With fifth generation and future wireless systems seeing significant increases in both network density and functionality, there are growing concerns over the environmental impact of cellular networks, and energy efficiency is becoming an important design consideration for physical layer technologies [230].

The high peak-to-average power ratio (PAPR) of OFDM waveforms, now ubiquitous in wireless systems, requires the use of linear power amplifiers (PAs) that operate with a large power back-off, resulting in poor power efficiency [101]. The power consumption of these PAs dominates BS RF power consumption, and PAPR reduction schemes that can improve efficiency by reducing PA power back-off, whilst preserving system performance, have long been a topic of academic and industrial research [231].

The shift to a massive MIMO paradigm has the potential to reduce overall radiated power and increase energy efficiency, whilst enabling the small number of high power, expensive, linear PAs used in a conventional BS to be replaced with a large number of low power, cheaper PAs [113]. However, the high PAPR of OFDM signals reduces the overall energy efficiency of these systems, making the use of reduced PAPR signals in massive MIMO systems desirable [149]. Fortunately, massive MIMO facilitates new means of addressing the OFDM PAPR problem through the exploitation of the large number of excess degrees of freedom at the transmitter, and has stimulated a new body of research into the problem.

This chapter investigates the adaptation of iterative clipping & filtering, a classical PAPR reduction technique, to exploit the degrees of freedom in the massive MIMO downlink for improved performance. It begins by developing a statistical model for the effect of non-linear signal distortion caused by clipping & filtering, using a vector Bussgang decomposition. This model shows that clipping & filtering has two effects on the MIMO downlink signal – it introduces additive clipping noise, and it attenuates and distorts the downlink symbol precoding. Numerical examples are provided to show the impact of clipping & filtering on link performance, providing motivation for the development of new methods that mitigate these effects.

In the second part of the chapter, a novel clipping and spatial filtering scheme is outlined that is capable of achieving high PAPR reduction with negligible loss in link performance. The scheme is based on generating, at each iteration, a least squares approximation of the clipped and filtered signal that is constrained to give a specified received signal at the users. The massive MIMO channel's large nullspace means that the error of this least squares approximation can be kept small, preserving the good PAPR properties of the clipped and filtered signal. It is shown that this least squares approximation can be generated as an additive combination of the clipped & filtered signal and a low power error cancellation signal that is precoded using the standard ZF matrix to correct the signal received at the users. The general architecture is shown in Figure 3.1.

Whilst least squares filtering-type approaches have previously been proposed, the work here improves on earlier work by explicitly accounting for the effects of the Bussgang clipping model when formulating the least squares problem (referred to here as Bussgang-aware least squares, BLS, filtering). This ensures that the algorithm converges at small clipping ratios, and enables greater PAPR reduction to be achieved. Numerical results for i.i.d & correlated Rayleigh fading channels show that for a typical massive MIMO configuration the scheme can achieve over 8 dB of PAPR reduction – 1 dB improvement on previous schemes – whilst incurring only 0.3 dB link performance degradation compared to ideal unclipped OFDM.

Finally, the BLS scheme is adapted to include active constellation extension (ACE), in which some distortion due to clipping is permitted providing it falls within certain regions of the symbol constellation. This is shown through numerical examples to provide an extra 1 dB of PAPR reduction in a massive MIMO setting when QPSK signalling is used, without any additional performance penalty. It is shown to be an attractive modification when the number of transmit antennas is reduced and when smaller QAM constellation sizes are used.

The proposed scheme relies on only linear processing, and has computational complexity that scales linearly with the MIMO dimensions, comparing favourably with the non-linear PAPR reduction techniques that have previously been proposed for massive MIMO. For typical MIMO configurations its overall complexity is around double that of conventional iterative clipping & filtering PAPR reduction. The scheme is able to achieve high levels of PAPR reduction with negligible loss in performance, and constitutes a promising practical method for PAPR reduction in massive MIMO systems.

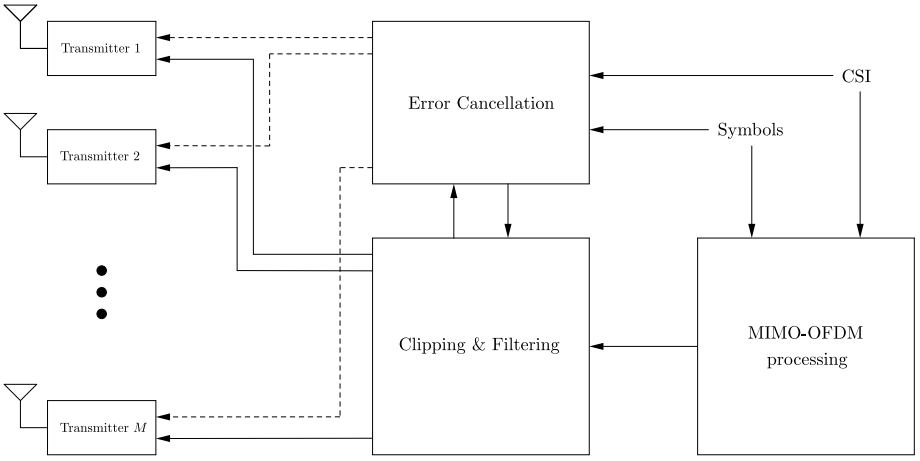

Figure 3.1: Block diagram of proposed PAPR reduction scheme.

3.1 Chapter Overview

The chapter has the following general structure:

- Section 3.2 provides the background to the PAPR problem in OFDM communication systems. The causes of and issues with high PAPR in OFDM are discussed, before a summary of classical approaches to PAPR reduction for SISO systems is given. The PAPR problem in MIMO systems is then discussed, and a review of state-of-the-art PAPR reduction methods for massive MIMO systems is provided.
- Section 3.3 derives a vector Bussgang model for analysing the effects of clipping & filtering in MIMO systems. It begins by analysing the effect of a single clipping & filtering iteration on the MIMO-OFDM signal, before generalising this to iterative clipping & filtering.
- Section 3.4 analyses the impact of standard clipping & filtering on massive MIMO performance using the Bussgang model, providing analysis and numerical results that demonstrate its limitations as a means of reducing PAPR.
- Section 3.5 develops the proposed Bussgang-aware least squares (BLS) PAPR reduction method. The limitations of prior work are discussed, before the proposed scheme is described – improving upon prior work by incorporating the Bussgang clipping model. The scheme is then adapted to incorporate active constellation extension for enhanced performance. Numerical results are provided throughout the chapter demonstrating its benefits.
- Section 3.6 summarises the findings and provides some concluding remarks and directions for future work.

3.1.1 Novel Contributions

The key contributions to the state-of-the-art made in this chapter are:

- **The derivation of a rigorous vector Bussgang model for analysing the effect of clipping on MIMO signals, sections 3.3.1, 3.3.2.** Whilst Bussgang's decomposition has been widely used to study non-linear effects in MIMO systems, and has previously been used for SISO clipping analysis, it has not previously been rigorously applied to the study of iterative clipping & filtering in MIMO systems. The development of this clipping model is foundational to the development of the BLS PAPR reduction scheme.

- **A clipping & spatial filtering PAPR reduction scheme that incorporates the Bussgang clipping model, sections 3.5.2, 3.5.3.** Previously proposed solutions have used a simple additive error model that does not adequately model the impact of non-linear clipping distortion on the MIMO signal, limiting their performance. Accounting for these effects enables the proposed scheme to achieve 8 dB of PAPR reduction – over 1 dB improvement compared to previous work.

- **The extension of the proposed PAPR reduction scheme to include active constellation extension, Section 3.5.5.** Despite being well established as a PAPR method for SISO OFDM systems, ACE has not been previously applied to the MU-MIMO downlink. Here it is shown that it is a natural companion to the spatial filtering approach, and can provide up to 1 dB of additional PAPR reduction in massive MIMO systems, with no extra performance degradation. When operating with a reduced number of BS transmit antennas, ACE can provide over 2 dB of additional PAPR reduction compared to just spatial filtering.

3.1.2 Published Work

The basis of the Bussgang model and clipping analysis in this work was published in [222]. However, the model used there made assumptions about the transmit signal and clipping noise statistics that do not always hold, and was applied to phase-only clipping, rather than iterative clipping & filtering. As such those assumptions have been removed here, with phase-only clipping given as a special case in Section 3.3.3.

Two of the main contributions of this chapter – the BLS PAPR reduction method and ACE method – are currently unpublished, due to time constraints. It is anticipated that another conference paper or letter could be published with these findings in.

3.2 Background

OFDM has become the de facto waveform of choice for modern high throughput wireless communication systems, due to its high spectral efficiency, flexibility and ease of use. Despite this ubiquity, however, the high peak-to-average power ratio (PAPR) of OFDM signals is a long standing problem, for which no universally applicable solution has been found [231]. A substantial body of research has therefore been produced, and continues to be produced, on the subject of PAPR reduction for OFDM.

This section begins by explaining the motivation for PAPR reduction, before outlining some of the most important and relevant solutions that have been proposed within the research

literature. Attention is first given to classical PAPR reduction techniques for SISO systems. It is then shown that the OFDM PAPR problem is exacerbated in MIMO systems, before PAPR reduction techniques that exploit specific opportunities provided by MIMO systems are discussed. A particular focus is given to methods related to clipping & filtering, since these techniques form the basis of the investigations and proposals of the new research in this chapter.

3.2.1 Motivation

The instantaneous transmit power of a wireless signal depends on the encoded information symbols, and is, in general, a continuously varying function of time,

$$p(t) = \left|x_{\mathrm{RF}}(t)\right|^2. \tag{3.1}$$

As discussed in Section 2.1.3, the capacity of a communication system is determined by the SNR or SINR of a link, which is a function of the *average* power of the transmit signal, $\mathbb{E}\big[p(t)\big]$. However, when designing a wireless system, the *peak* instantaneous power is also a crucial parameter, since this determines the power level that the power amplifier (PA) in the transmitter must be capable of supplying.

An important characteristic of a signal is therefore its peak-to-average power ratio

$$\mathrm{PAPR} = \frac{\text{peak power}}{\text{average power}}. \tag{3.2}$$

Power amplifiers are inherently non-linear devices that have an input-output characteristic that is only approximately linear below a certain input power level. Operating beyond this level results in non-linear distortion of the input signal, causing performance degradation and spurious out-of-band emissions. The input must therefore be 'backed-off' to ensure that the PA operates within its linear region at all power levels, as shown in Figure 3.2.

This backoff is determined by the PAPR of the transmit signal – a high PAPR requiring a large power backoff. The use of signals with high PAPR therefore has a number of disadvantages:

- For a given performance level (average power) the PA must be capable of supplying a higher peak power, increasing the cost and size of the power amplifier. Alternatively, for a given PA peak power a lower average power must be used, resulting in worse performance.

- Power amplifiers generally operate at maximum efficiency at their peak power level, and therefore the use of large power backoff results in poor power efficiency. This can be illustrated (somewhat simplistically) by considering an ideal class A PA, which has constant power consumption and efficiency of 50% at peak power,

  $$\text{power efficiency} = \frac{0.5}{\text{power backoff}}. \tag{3.3}$$

  For a maximum expected PAPR of 10 dB, a power backoff of 10 dB will be used, resulting in an efficiency of just 5%. Modern PA architectures can improve this efficiency – at the expense of increased complexity and cost – but also operate most efficiently with small backoffs [25].

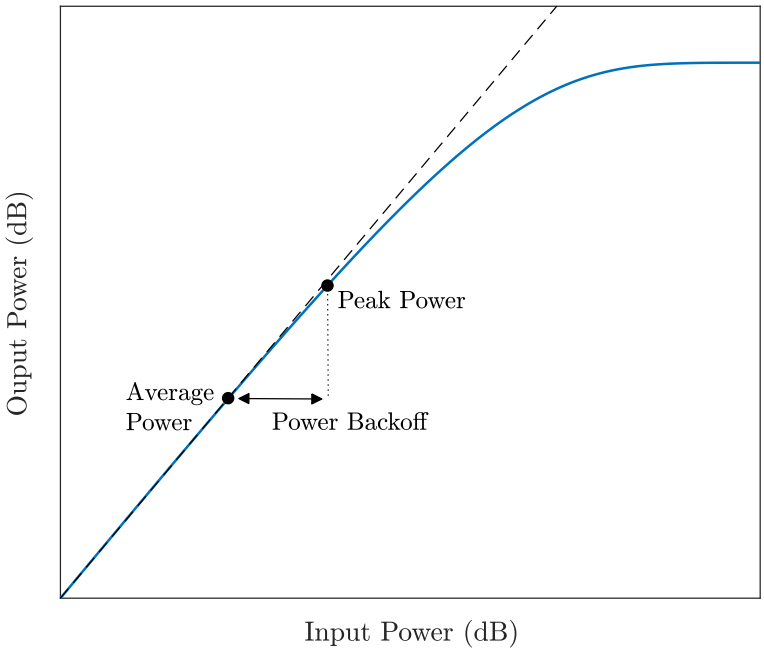

Figure 3.2: Example power amplifier input-output characteristic.

- The digital transmit signal samples must be converted to analogue using a digital to analogue convertor (DAC). A larger PAPR & backoff means higher resolution DACs must be used, which have higher power consumption & cost [190].

The PAPR of the RF transmit signal, $x_{\text{RF}}(t)$, is given by

$$\text{RF PAPR} = \frac{\max\limits_{t} p(t)}{\mathbb{E}\big[p(t)\big]}. \tag{3.4}$$

Due to the presence of the RF carrier, this is approximately double the PAPR of the baseband transmit signal [101]

$$\text{RF PAPR} \approx 2 \times \text{baseband PAPR}, \tag{3.5}$$

which is similarly defined,

$$\text{baseband PAPR} = \frac{\max\limits_{t} \big|x(t)\big|^2}{\mathbb{E}\Big[\big|x(t)\big|^2\Big]}. \tag{3.6}$$

In practice, it is more convenient to measure PAPR in the digital domain. Assuming appropriate (e.g sinc) pulse shaping is used, the baseband PAPR can be estimated from the digital baseband transmit signal as

$$\text{baseband PAPR} \approx \frac{\max\limits_{n} \big|x[n]\big|^2}{\mathbb{E}\Big[\big|x[n]\big|^2\Big]}, \tag{3.7}$$

where n represents the sample number, and it is assumed that the transmit signal, $x[n]$, is over-sampled by a factor of q in order to accurately capture the peaks in the signal. An oversampling

factor of $q \geq 4$ is sufficient for accurate results [101]. Herein, the term PAPR refers to baseband PAPR (noting that this has a direct relationship with RF PAPR).

OFDM is ubiquitous in modern high performance wireless applications due to its ability to deliver high spectral efficiencies, and is the waveform of choice in fifth generation and future WiFi standards. However, its time domain signal, which is the aggregate of many independently modulated subcarriers, naturally suffers from a high PAPR.

The oversampled SISO OFDM samples, $x[n]$, are given by

$$x[n] = \frac{1}{\sqrt{N}} \sum_{l=0}^{N-1} s_l e^{j2\pi \frac{ln}{qN}}. \tag{3.8}$$

If the chosen symbol alphabet (typically a QAM constellation) has average power $\mathbb{E}\left[|s_l|^2\right] = \rho$, with peak symbol power $\rho_{\max}$, then the maximum possible PAPR of an OFDM symbols is $N\frac{\rho_{\max}}{\rho}$ (occurring if the same peak power symbol is transmitted simultaneously on every subcarrier). As N grows large the maximum PAPR becomes very large, but the probability of this peak power occurring also becomes very small. For large N, by the central limit theorem the time domain samples tend towards following a complex normal distribution,

$$x[n] \sim \mathcal{CN}(0, \rho). \tag{3.9}$$

The PAPR of each OFDM symbol depends on the random information message symbols, and therefore itself randomly varies between symbols [101], as shown by the complementary cumulative distribution function in Figure 3.3. The 'PAPR of a signal' can be quoted as the maximum

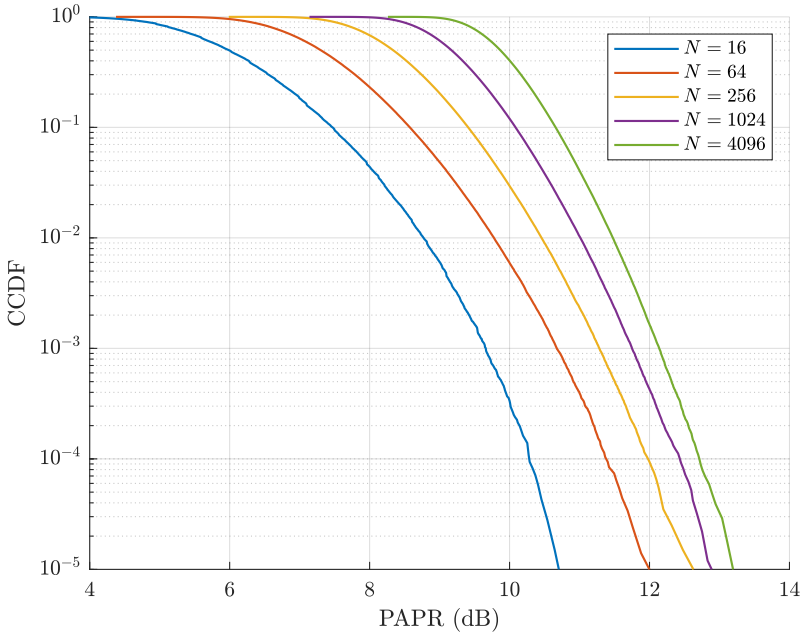

Figure 3.3: Complementary cumulative distribution function of OFDM PAPR (baseband), varying number of subcarriers (N), QPSK symbols.

PAPR value that is exceeded by less than 1 in 1,000 symbols (other figures, e.g. 1 in 10,000

may also be used). Figure 3.3 shows that OFDM symbols typically have a maximum expected PAPR well in excess of 10 dB, and the use of PAPR reduction techniques is therefore vital to improving PA efficiency and reducing cost.

3.2.2 Classical PAPR Reduction

A large variety of techniques for reducing the PAPR in SISO OFDM systems have been proposed, each having associated benefits and costs in terms of PAPR reduction capability, link performance degradation, signalling overheads and computational complexity.

Clipping & Filtering

The simplest – and mostly widely used – PAPR reduction method is clipping, in which the peak amplitude of the oversampled signal is digitally limited prior to analogue conversion [158]. This can in principle achieve an arbitrary PAPR reduction, but at the expense of introducing distortion, or *clipping noise*, into the signal. This clipping noise is spread over all frequencies, causing both in-band and out-of-band distortion, and increases in power when clipping is applied more aggressively.

The out-of-band distortion from clipping can violate out-of-band emission regulations, and clipping must therefore be followed by filtering to remove these signal components [158]. This filtering operation causes signal peak regrowth, and so clipping & filtering is generally applied repeatedly, until convergence to a signal with reduced PAPR [6]. A number of schemes have sought to address the increased computational complexity of repeated clipping & filtering, e.g [127].

The filtering operation does not alleviate the effects of in-band distortion, which introduces a noise-like error on each subcarrier, degrading the link performance. Methods for estimating and cancelling in-band distortion at the receiver have therefore been proposed, e.g. [78]. For improved performance, clipping & filtering can also be applied in conjunction with other PAPR reduction techniques.

Active Constellation Extension

The active constellation extension (ACE) method is based on the observation that, for finite, discrete symbol constellations – such as QAM – the outermost constellation points can be extended infinitely without reducing the probability of symbols being accurately decoded [112]. Transmit symbols that occupy these outermost constellation points can therefore be dynamically extended in order to reduce PAPR. This is illustrated for QPSK, where the constellation points can be extended into shaded regions without reducing the minimum constellation spacing, $d_{\min}$.

The PAPR minimising ACE transmit symbols on each subcarrier can be found as the solution to a convex optimisation problem. However, a more practical but sub-optimal solution is to use clipping & filtering to generate a reduced PAPR signal, and then modify the resulting distorted symbols so that they lie at the nearest point within the appropriate constellation extension region, repeating iteratively until convergence. Gradient based methods to improve the convergence of the algorithm have been proposed in [112] and [8].

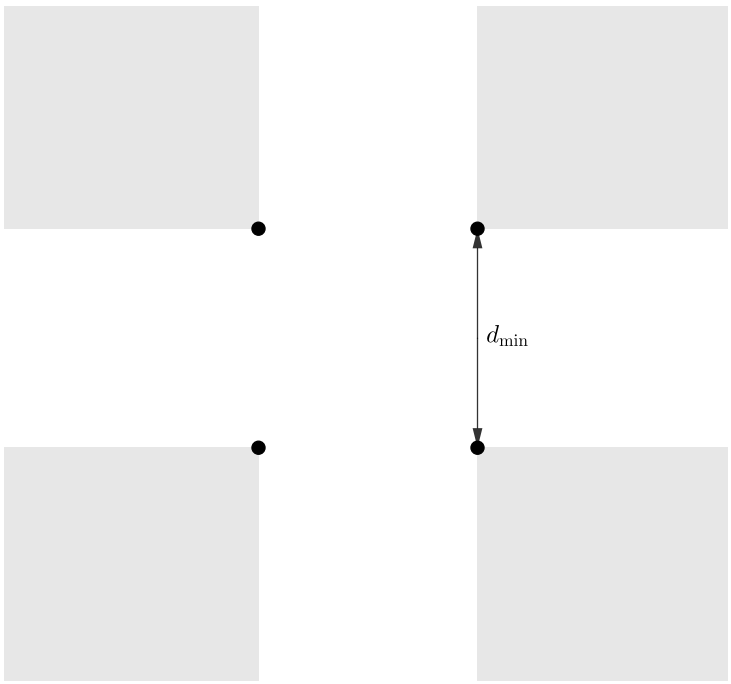

Figure 3.4: Active constellation extension regions (shaded) for QPSK.

ACE has the unique property of imposing no link performance penalty whilst also requiring no additional side information at the receivers – the use of ACE at the transmitter is transparent to the receiver. A modification to the ACE method to further reduce PAPR by allowing a small amount of distortion has been proposed in [186].

Partial Transmit Sequences & Selected Mapping

Another approach is to modulate sub-blocks of subcarriers and then combine them in a such way that the overall OFDM symbol produced has low PAPR.

In the partial transmit sequence method, N_{PTS} blocks of N/N_{PTS} subcarriers are modulated separately to produce N_{PTS} partial transmit sequences. An optimised phase weighting is then applied to each sequence and they are combined to form a full OFDM symbol, with the phase weightings chosen to minimise PAPR . These phase weightings must be communicated to the receiver as side information, with overheads kept low by restricting the set of possible phase weightings [150]. To reduce computational complexity, various sub-optimal methods for choosing the phase weights have been proposed, e.g. [42].

Closely related is the selected mapping method, in which a selection of candidate OFDM symbols are generating using pre-determined sets of phase weightings (applied to either individual or sub-blocks of subcarriers), and the symbol with the lowest PAPR selected [13]. This has the benefit of reduced side information overhead, since only the index of the selected set of phase weights needs to be communicated.

Other Methods

Other methods for PAPR reduction include [101]:

- Tone injection – the discrete symbol constellation lattice is repeated infinitely, such that each of the original constellation points maps to many points. The transmit points on the new lattice that (approximately) minimise PAPR are then found, with the original signal recovered at the receiver using a modulo operation.
- Tone reservation – certain subcarriers are reserved from carrying data, and used to instead transmit arbitrary symbols that are chosen to cancel the peaks in the data signal.
- Companding – similar to clipping, except that an invertible, soft clipping function is used instead of hard limiting. After equalisation at the receiver, the original signal can then be approximately recovered.

3.2.3 PAPR Reduction in MIMO Systems

As discussed in Section 2.2.1, OFDM is the natural choice of waveform in MIMO systems, converting the high dimensional multi-antenna time domain channel into a set of parallel vector multiplications.

Consider a linearly precoded MIMO-OFDM signal,

$$\mathbf{x}[n] = \frac{1}{\sqrt{N}} \sum_{l=0}^{N-1} \mathbf{x}_l e^{j2\pi \frac{ln}{qN}} \tag{3.10}$$

where

$$\mathbf{x}_l = \mathbf{W}_l \mathbf{s}_l, \tag{3.11}$$

where $\mathbf{W}_l$ is the precoding matrix on subcarrier l (e.g. ZF or MMSE). By the central limit theorem, the MIMO-OFDM signal samples can be modelled, for large N, as coming from a wide sense stationary complex normal distribution,

$$\mathbf{x}[n] \sim \mathcal{CN}\big(0, \mathbf{R}_x\big) \tag{3.12}$$

where

$$\mathbf{R}_x = \mathbb{E}\big[\mathbf{x}[n]\mathbf{x}^\dagger[n]\big] \tag{3.13}$$

$$= \frac{1}{N} \sum_{l=0}^{N-1} \mathbf{W}_l \mathbf{W}_l^\dagger. \tag{3.14}$$

The average total transmit power of the system is

$$P_T = \mathrm{Tr}\left(\mathbf{R}_x\right), \tag{3.15}$$

whilst the average transmit power of antenna m is

$$P_m = \big[\mathbf{R}_x\big]_{m,m}. \tag{3.16}$$

On a per-antenna level, the precoded OFDM signals follow the same PAPR distribution as in SISO systems. However, since the power must be backed off equally at all antennas, the required

backoff is determined by the antenna with the peak power. An 'array PAPR' metric is thus a more appropriate measure of the PAPR performance of any scheme

$$\text{array PAPR} = \frac{\text{peak per-antenna power}}{\text{average per-antenna power}} \tag{3.17}$$

$$= \frac{\max\limits_{m,n} \left|x_m[n]\right|^2}{P_T/M}. \tag{3.18}$$

The array PAPR is necessarily higher than the per-antenna PAPR, for two reasons:

- The increased number of signal samples means that the probability of a certain peak power value being reached within one OFDM symbol is greater. This is illustrated in Figure 3.5 for a MIMO system with 8 users and $N = 512$ subcarriers (where each antenna experiences independent fading with 10 Rayleigh distributed time domain channel taps).

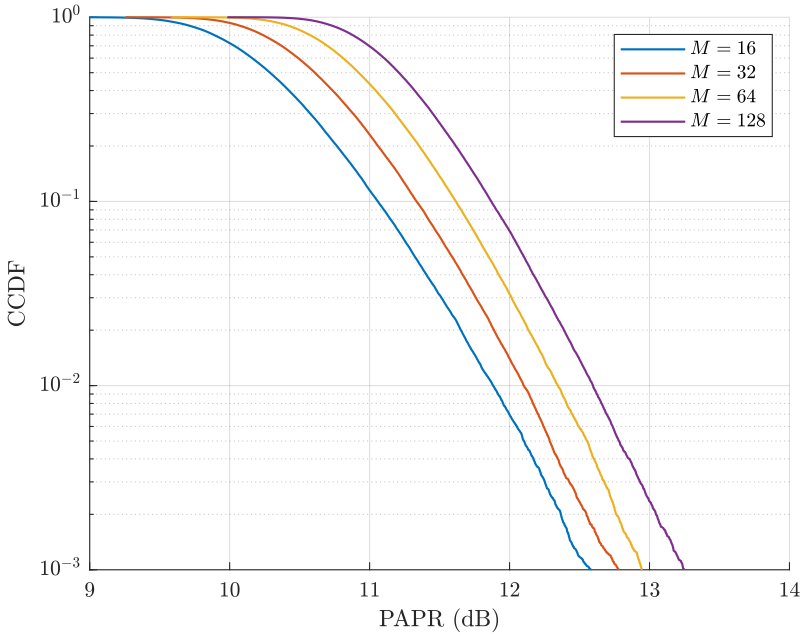

Figure 3.5: Complementary cumulative distribution function of MIMO-OFDM PAPR, varying number of BS antennas (M), 10 time domain i.i.d Rayleigh fading channel taps, 512 subcarriers, QPSK symbols.

- The use of spatial precoding means that the average and peak transmit powers may differ between transmit antennas. Conventional precoding methods as outlined in Section 2.3 optimise performance subject to a total power constraint (TPC), and can result in significant variation in per-antenna average power, particularly in spatially correlated channels. This is illustrated in Figure 3.6 for a correlated Rayleigh fading channel with a $M = 64$ element horizontal uniform linear array, 10° azimuth spread and $K = 8$ users roughly equally spaced, equidistant from the BS, within a single sector of a three sector cell. Precoding under per-antenna power constraints (PAPC), using the method in [242], reduces PAPR in the correlated fading channel by 2.5 dB, but is much more computationally expensive.

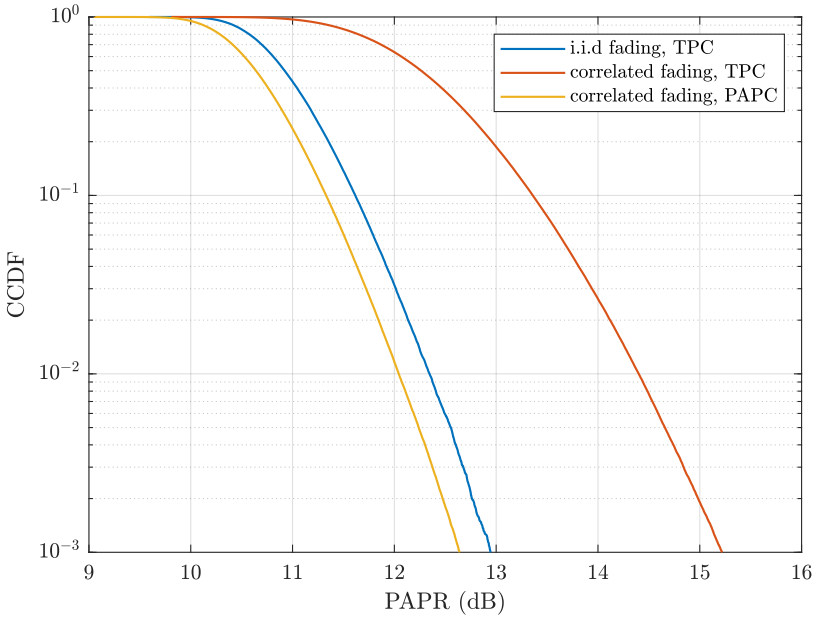

Figure 3.6: Complementary cumulative distribution function of MIMO-OFDM PAPR in correlated channel, with different precodings.

The increased PAPR indicates that PAs in massive MIMO systems will tend to have to operate with a larger backoff than in SISO systems, reducing their efficiency. Fortunately, some properties of massive MIMO help mitigate this problem.

Firstly, it is known from Chapter 2 that the array gain of a massive MIMO system can significantly reduce overall, so that overall energy efficiency may still be improved despite a decrease in individual PA efficiency [16].

Secondly, it is shown in [147] that both in-band and out-of-band non-linear PA distortion are beamformed with a smaller gain than the precoded information symbols, and thus their impact reduces. This suggests that PAs may be allowed to operate beyond their linear operating region without necessarily violating spectral regulations or significantly impacting link performance.

Finally – the focus of this chapter – the massive MIMO channel facilitates new ways of reducing the PAPR of transmit signals without incurring significant link performance loss, thanks to its large nullspace. The key feature that enables these methods is that on each subcarrier, an additional signal $\mathbf{e} \in \mathbb{C}^M$ may be transmitted,

$$\mathbf{x}_l = \mathbf{W}_l \mathbf{s}_l + \mathbf{e}, \tag{3.19}$$

without having any effect on the received signal,

$$\mathbf{H}_l \mathbf{x}_l = \mathbf{H}_l \mathbf{W}_l \mathbf{s}_l, \tag{3.20}$$

providing $\mathbf{e}$ lies in the channel nullspace, i.e.

$$\mathbf{H}_l \mathbf{e} = 0. \tag{3.21}$$

This nullspace has dimension $M - K$, and thus with a large excess of BS antennas a huge (infinite) range of different $\mathbf{e}$ vectors can be inserted onto each subcarrier without introducing any additional interference at the K users. The PAPR of the MIMO-OFDM signal may therefore be reduced by using a peak reduction/cancellation signal that lies (or approximately lies) within the channel nullspace – at the expense of increased average transmit power due to the additional power associated with $\mathbf{e}$.

Nullspace-based PAPR reduction has now received significant research attention. Some of the proposed methods use it explicitly, by performing standard massive MIMO-OFDM precoding and then finding a PAPR reducing signal, and others implicitly – directly choosing the $\mathbf{x}_l$ to give the desired response at the users whilst minimising transmit PAPR. The former methods often have the benefit of being add-ons to a standard MIMO-OFDM PHY architecture, whilst the latter category benefit from having potentially better performance, at the expense of greater computational complexity due to their need for each symbol to precoded individually.

A review of research into these methods is now provided, with a particular focus on techniques which employ clipping – the basis of the techniques proposed in this chapter. Alternative PAPR reduction methods that do not exploit the characteristics of the massive MIMO channel have also been proposed, but are outside the focus of this work.

Constant Envelope & Low PAPR Precoding

In [142] it is shown that for narrowband single carrier massive MIMO systems, it is possible to use 'constant envelope' (CE) signals, constrained to the form

$$x_m[n] = \gamma e^{j\theta_m[n]} \tag{3.22}$$

whilst, for sufficiently large M, still suppressing inter-user interference and achieving an array gain. A non-linear least squares optimisation for finding the phase angles, $\theta_m[n]$, is given. A number of follow up works propose alternative narrowband CE precoding schemes for massive MIMO systems, such as [4] and [34], and have extended this idea to more realistic wideband channels in [141] and [119].

However, it is important to note that CE precoding schemes are not truly constant envelope, as they generally only have a phase-only characteristic when critically sampled at the Nyquist frequency, and therefore their time domain signals do not have a constant power level. The stringent constant envelope condition in (3.22) therefore does not necessarily produce the precoded signals with the overall lowest PAPR. A true (continuous time) constant envelope precoding scheme is derived in [148], but exhibits poor spectral efficiency.

In [203] the authors instead formulate a convex optimisation problem that minimises the peak transmit power, whilst giving a desired received signal pattern. The proposed 'FITRA' algorithm is able to approximate any linear precoder and is implemented by a simple iterative algorithm. It is seen to outperform conventional clipping at low PAPR values, but the authors note that due to the large number of iterations required for convergence it has complexity 'one-to-two orders of magnitude larger' than linear precoding. Various other PAPR-aware precoding methods have been proposed, e.g. [9], [33], [239]. However, since precoding is performed

iteratively and on a per-symbol basis (in contrast to linear precoders, which have closed form solutions that only need to be calculated once per coherence block), they incur a significant increase in computational complexity.

Clipping & Spatial Filtering Methods

Conventional clipping methods use frequency domain filtering to eliminate the out-of-band effects of clipping, but do nothing to mitigate the deleterious effects of clipping on in-band performance. The addition of a *spatial* filtering stage can be seen as a generalisation of the clipping & filtering scheme for MIMO systems, that enables the in-band error at the users to be eliminated. This idea was first applied to single user MIMO in [205], where a reduced number of data streams are transmitted and the clipping-induced error filtered to lie only in the remaining channel nullspace, such that no link degradation occurs.

This idea is naturally extended to massive MIMO systems, where a large channel nullspace is inherently present, as in [206]. There, filtering is achieved by using the singular value decomposition to calculate $M - K$ basis vectors for the channel nullspace, and then projecting the clipping error (the difference between the clipped signal and original precoded OFDM signal) into this nullspace. When M/K is large, much of the error often naturally lies within the nullspace, and hence the difference between the two signals is small.

As observed in [219], the nullspace projection method, applied on each subcarrier, gives a least squares approximation of the clipped signal – minimising the squared difference between the transmit signal and the clipped signal, whilst giving the desired response at the users. Whilst this least squares approximation does not explicitly limit PAPR – which depends on the maximum peak power, or ℓ_∞ norm – it acts as an effective and simple proxy, particularly when applied iteratively, and has been shown to produce low PAPR signals using only linear processing. A more complex algorithm is proposed in [243], where symbol precoding is also incorporated into the iterative clipping and nullspace projection algorithm, at the expense of increased computational cost.

An alternative approach is taken in [172], where a heavily clipped OFDM signal is transmitted using the majority of the BS antennas, with the remaining antennas reserved to transmit a low power precoded signal that cancels out the error caused by clipping at the users. The research later in this chapter shows that this method is in fact closely related to the clipping & spatial filtering approach, which can be seen as transmitting both the clipped OFDM signal plus a small cancellation signal simultaneously with *all* antennas.

Other PAPR Reduction Methods

Other approaches have been taken to finding nullspace-based peak cancellation signals, such as [31], where the cancellation signal inserted onto each subcarrier is optimised, with the ability to trade-off link performance against PAPR reduction. The proposed algorithm significantly reduces PAPR, but requires a numerical optimisation with complexity $\geq \mathcal{O}(N^3)$ to be performed for each OFDM symbol. In [107] a set of candidate cancellation signals are generated randomly and the best one selected, with modest results.

Methods that do not exploit the large nullspace of the MIMO channel include an adaptation of the partial transmit sequence scheme for MIMO systems [109], and a tone reservation scheme that concentrates on peak cancellation only on the antennas experiencing highest peak power [156].

3.3 Clipping & Filtering in Massive MIMO-OFDM Systems

This chapter develops an analytically tractable statistical model for the non-linear distortion that occurs when clipping & filtering is applied to a MIMO-OFDM signal, using Bussgang's theorem.

3.3.1 Clipping & Filtering of MIMO-OFDM

In a MIMO-OFDM downlink system, clipping is applied to each of the M antenna signals to limit the peak signal magnitude at each antenna to a maximum level, $A_{\max}$. The value of $A_{\max}$ is fixed across all antennas, in order to impose a uniform peak power limit, and usefully written as a function of an 'array clipping ratio', γ and the average per-antenna power, P_T/M,

$$A_{\max} = \gamma\sqrt{P_T/M}, \tag{3.23}$$

where $P_T = \mathrm{Tr}\left(\mathbf{R}_x\right)$. The clipping operation is a defined by a non-linear function,

$$x_{\mathrm{C},m}[n] = f_{\mathrm{C}}\big(x_m[n], A_{\max}\big), \tag{3.24}$$

given by

$$f_{\mathrm{C}}(x, A_{\max}) = \begin{cases} x, & |x| \leq A_{\max} \\ A_{\max}\dfrac{x}{|x|}, & |x| > A_{\max}. \end{cases} \tag{3.25}$$

The peak clipped signal power is $\gamma^2 P_T/M$, and it is straightforward to see from (3.17) that the PAPR of the clipped signal is then approximately[1]

$$\text{array PAPR} \approx \gamma^2. \tag{3.26}$$

In previous work [105], an additive error model has been used to model the effect of clipping

$$\mathbf{x}_{\mathrm{C}}[n] = \mathbf{x}[n] + \mathbf{c}[n]. \tag{3.27}$$

However, the additive error $\mathbf{c}[n]$ is correlated with the original OFDM signal[2], $\mathbf{x}[n]$, limiting its usefulness for analysing the effects of clipping.

Fortunately, since the OFDM signals are Gaussian distributed, Bussgang's theorem [26] can be applied to decompose the clipped signal into two uncorrelated components, a technique that

[1] Since the clipping operation decreases the average as well as the peak power of the signal, this approximation doesn't hold for small clipping ratios, where a significant proportion of samples are clipped and average power reduced significantly. However, the relationship between clipping ratio and PAPR remains monotonic.

[2] This is readily inferred from the fact that clipping reduces signal power, and hence the power of the clipped signal must be less than the original OFDM signal, $\mathbb{E}\big[\|\mathbf{x}[n] + \mathbf{c}[n]\|^2\big] < \mathbb{E}\big[\|\mathbf{x}[n]\|^2\big]$

has previously been used to analyse the effect of clipping in SISO OFDM systems [158]. The Bussgang decomposition is

$$\mathbf{x}_{\mathrm{C},m}[n] = \mathbf{A}\mathbf{x}[n] + \boldsymbol{\varepsilon}[n] \tag{3.28}$$

where $\mathbf{A}$ is the Bussgang gain and $\boldsymbol{\varepsilon}[n]$ is *clipping noise* that is uncorrelated with the original OFDM signal

$$\mathbb{E}\big[\mathbf{x}[n]\boldsymbol{\varepsilon}^{\dagger}[n]\big] = \mathbf{0}. \tag{3.29}$$

The Bussgang gain is a diagonal scaling matrix $\mathbf{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_M)$ [50], with elements $0 < \alpha_m < 1$ given by [158]

$$\alpha_m = \frac{\mathbb{E}\big[x_{\mathrm{C},m}[n]x_m^*[n]\big]}{\mathbb{E}\big[\big|x_m[n]\big|^2\big]} = \frac{\mathbb{E}\big[x_{\mathrm{C},m}[n]x_m^*[n]\big]}{P_m} \tag{3.30}$$

$$= 1 - e^{-\gamma_m^2} + \frac{\sqrt{\pi}\gamma_m}{2}\mathrm{erfc}(\gamma_m), \tag{3.31}$$

a monotonically decreasing function of γ_m (the per-antenna clipping ratio at antenna m)

$$\gamma_m = \frac{A_{\max}}{\sqrt{P_m}} = \gamma\frac{\sqrt{P_T/M}}{\sqrt{P_m}}. \tag{3.32}$$

The Bussgang model reveals that clipping has two effects – it transforms the OFDM signal component, and it introduces clipping noise into the transmission. When each antenna has the same average power the Bussgang gain is of the form $\alpha\mathbf{I}_M$, and the OFDM signal component is simply attenuated by a factor α. For the general case of differing per-antenna powers, this scaling effectively distorts the OFDM component by attenuating the signal at each antenna by differing amounts. The power of the clipping noise can be quantified as in [158], but it suffices here to note that it increases in power relative to the OFDM signal as γ is decreased.

The Bussgang model is stationary whilst the statistics of $\mathbf{x}[n]$ are stationary – i.e. whilst the set of linear precoders is fixed. The Bussgain gain also remains constant under a scaling in transmit power, providing there is a corresponding scaling in the clipping level, $A_{\max}$ (such that γ remains constant), with the clipping noise power scaling accordingly. The use of a Bussgang signal model here is a key improvement that facilitates the algorithm design and analysis in this research chapter.

In the frequency domain, by the linearity of the Fourier transform, the signal on each subcarrier is

$$\mathbf{x}_{\mathrm{C},l} = \begin{cases} \mathbf{A}\mathbf{x}_l + \boldsymbol{\varepsilon}_l & l \in [0, N-1] \\ \boldsymbol{\varepsilon}_l & l \in [N, qN-1]. \end{cases} \tag{3.33}$$

The in-band subcarriers contain a transformed version of the precoded MIMO signal, corrupted by clipping noise, whilst the out-of-band subcarriers contain only clipping noise. This frequency domain clipping noise is given by

$$\boldsymbol{\varepsilon}_l = \frac{1}{N}\sum_{n=0}^{qN-1} \boldsymbol{\varepsilon}[n]e^{-j2\pi\frac{ln}{qN}}, \tag{3.34}$$

and, by the central limit theorem, is approximately jointly Gaussian,

$$\boldsymbol{\varepsilon}_l \sim \mathcal{CN}\big(0, \mathbf{R}_{\varepsilon,l}\big), \tag{3.35}$$

as well as being uncorrelated with the data signal,

$$\mathbb{E}\big[\mathbf{x}_l \boldsymbol{\varepsilon}_l^{\dagger}\big] = \mathbf{0}. \tag{3.36}$$

The covariance matrix,

$$\mathbf{R}_{\varepsilon,l} = \mathbb{E}\big[\boldsymbol{\varepsilon}_l \boldsymbol{\varepsilon}_l^{\dagger}\big], \tag{3.37}$$

describes the spatial characteristics of the clipping noise. This varies with frequency, l (as can be inferred from the analysis in [158]), but – due to its non-linear relationship with $\mathbf{x}[n]$ – a convenient closed form expression is not available, and hence it must be estimated numerically.

The out-of-band clipping noise components in $\mathbf{x}_{\mathrm{C}}[n]$ must be filtered for the transmission signal to comply with regulatory licenses. Perfect filtering of these components is achieved by simply setting the out-of-band subcarriers to zero,

$$\mathbf{x}_{\mathrm{CF},l} = \begin{cases} \mathbf{x}_{\mathrm{C},l} & l \in [0, N-1] \\ \mathbf{0} & l \in [N, qN-1]. \end{cases}, \tag{3.38}$$

leading to the clipped and filtered OFDM signal,

$$\mathbf{x}_{\mathrm{CF}}[n] = \frac{1}{N} \sum_{l=0}^{N-1} \mathbf{x}_{\mathrm{CF},l} e^{-j2\pi \frac{ln}{qN}}. \tag{3.39}$$

The frequency domain filtering operation causes regrowth of some signal peaks, and hence the PAPR of the clipped and filtered signal is considerably higher than indicated by (3.26). The clipping operation has computational complexity $\mathcal{O}\big(qMN\big)$ and hence complexity is dominated by the FFT and IFFT required for the filtering operation, with overall complexity $\mathcal{O}\big(qMN \log_2(qN)\big)$.

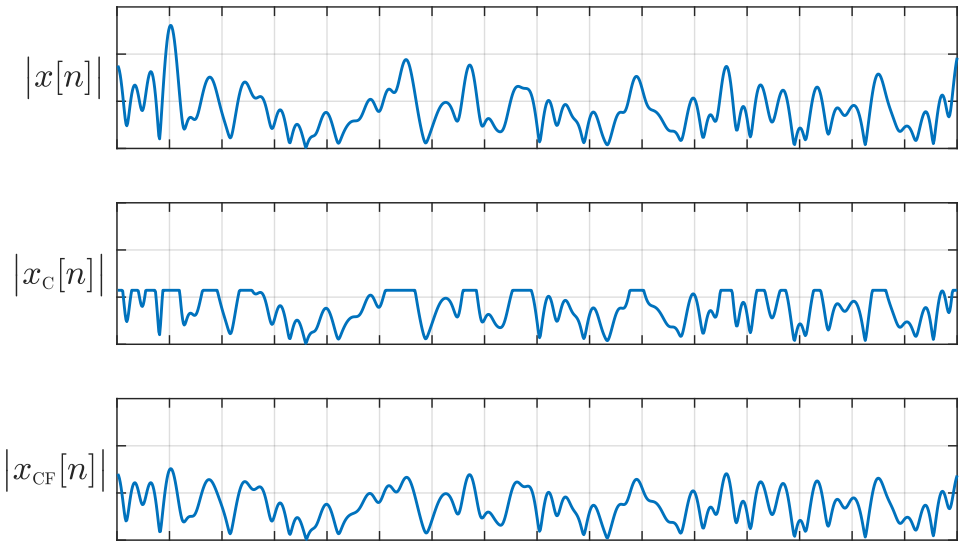

Figure 3.7: Original OFDM signal (top), clipped OFDM signal (middle), clipped & filtered OFDM signal (bottom).

3.3.2 Iterative Clipping & Filtering

To combat the peak regrowth it is common to apply clipping & filtering iteratively,

$$x_{\mathrm{C},m}^{(i+1)}[n] = f_{\mathrm{C}}\big(x_{\mathrm{CF},m}^{(i)}[n], A\big) \tag{3.40}$$

until the signal converges to a reduced PAPR, as shown in Figure 3.8.

As the clipped and filtered input to the clipping function is no longer Gaussian distributed, Bussgang's theorem cannot be directly applied. However, a statistically equivalent model of the clipped signal still exists in the same form,

$$\mathbf{x}_{\mathrm{C}}^{(i)}[n] = \mathbf{A}^{(i)}\mathbf{x}_{\mathrm{CF}}^{(i-1)}[n] + \boldsymbol{\varepsilon}^{(i)}[n] \tag{3.41}$$

where the clipping noise is uncorrelated with the clipped and filtered signal and $\mathbf{A}^{(i)}$ is given by [50]

$$\mathbf{A}^{(i)} = \mathbf{R}_{x_c,x_{cf}}^{(i)}\left(\mathbf{R}_{x_{cf}}^{(i-1)}\right)^{-1}, \tag{3.42}$$

with

$$\mathbf{R}_{x_c,x_{cf}}^{(i)} = \mathbb{E}\big[\mathbf{x}_{\mathrm{C}}^{(i)}[n]\mathbf{x}_{\mathrm{CF}}^{(i-1)\dagger}[n]\big], \tag{3.43}$$

$$\mathbf{R}_{x_{cf}}^{(i-1)} = \mathbb{E}\big[\mathbf{x}_{\mathrm{CF}}^{(i-1)}[n]\mathbf{x}_{\mathrm{CF}}^{(i-1)\dagger}[n]\big]. \tag{3.44}$$

The final clipped and filtered signal can be expressed,

$$\mathbf{x}_{\mathrm{CF}}[n] = \mathbf{A}\mathbf{x}[n] + \boldsymbol{\varepsilon}[n] \tag{3.45}$$

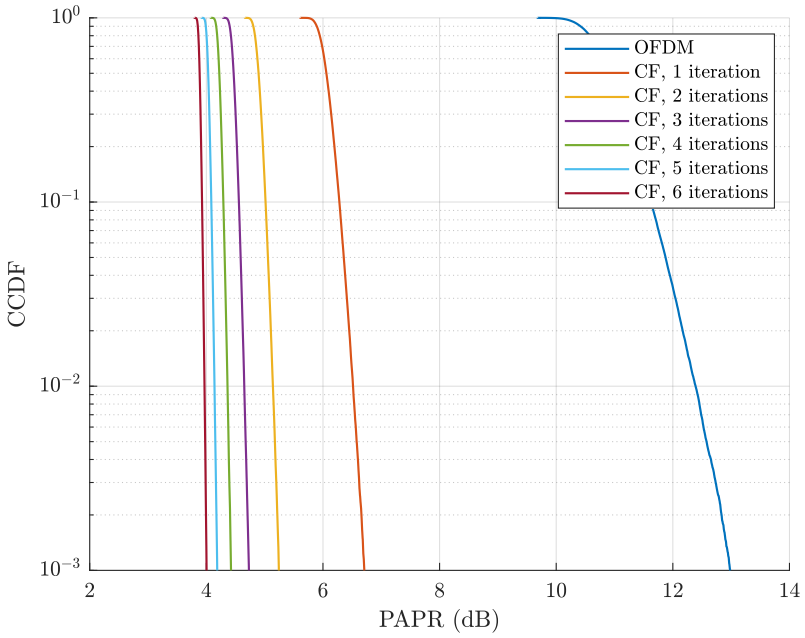

Figure 3.8: Complementary cumulative distribution function of array PAPR with iterative clipping & filtering, $\gamma = 1.2$, 8 users, 64 BS antennas, 512 subcarriers, QPSK symbols.

where

$$\mathbb{E}\big[\mathbf{x}[n]\boldsymbol{\varepsilon}^{\dagger}[n]\big] = \mathbf{0}, \tag{3.46}$$

and

$$\mathbf{A} = \prod_{i} \mathbf{A}^{(i)} \tag{3.47}$$

$$= \mathbf{R}_{x_{cf},x}\mathbf{R}_{x}^{-1} \tag{3.48}$$

with

$$\mathbf{R}_{x_{cf},x} = \mathbb{E}\big[\mathbf{x}_{\mathrm{CF}}[n]\mathbf{x}^{\dagger}[n]\big]. \tag{3.49}$$

As with the Bussgang decomposition, on a subcarrier level the signal is similarly expressed,

$$\mathbf{x}_{\mathrm{CF},l} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}_{l}. \tag{3.50}$$

This is referred to here as the generalised Bussgang model, and can be analysed in the same way as the Bussgang model. The main difference is that the matrix $\mathbf{A}$ is now generally non-diagonal, and must be estimated numerically. The exception here is the case where the transmit samples are independent at each antenna, in which case $\mathbf{R}_{x_{cf},x}$ and $\mathbf{R}_{x}$ are both diagonal matrices, resulting in diagonal $\mathbf{A}$ and $\mathbf{R}_{\varepsilon}$.

3.3.3 Phase-only OFDM

The ideal transmit signal has a PAPR of 1 – all samples have the same amplitude. This can be produced by a 'phase-only' clipping function in which all samples are clipped to a constant

amplitude, A_{P}, with only the phase information of the OFDM signal preserved,

$$f_{\mathrm{P}}(x) = A_{\mathrm{P}} \frac{x}{|x|}. \tag{3.51}$$

As with standard clipping & filtering, the phase-only signal can be represented using the Bussgang decomposition, with

$$\alpha_m = \frac{\mathbb{E}\big[|x_m[n]|\big]}{\mathbb{E}\big[|x_m[n]|^2\big]} = \frac{A_{\mathrm{P}}}{2}\sqrt{\frac{\pi}{P_m}}, \tag{3.52}$$

which follows directly from properties of the Rayleigh distribution. Thus the Bussgang gain for each antenna varies inversely with the per-antenna power – high power antennas experiencing greater signal attenuation. This Bussgang decomposition gives an interesting insight – just the phase information of the OFDM signal is sufficient to communicate some information [158].

Since the necessary filtering of out-of-band signal components causes peak regrowth, as in 3.2.2, improved performance is achieved by applying the phase-only and filtering scheme (PF) iteratively. Whilst the final PF signal does not have a constant envelope, it can be considered 'phase-only' in the sense that only the phase information from the original OFDM signal is present. The analysis in the following sections can be applied equally to clipped & filtered OFDM or phase-only OFDM (which can be seen as a special case of CF).

3.4 Impact of Clipping & Filtering on Massive MIMO Performance

The received clipped and filtered signal on a given subcarrier is given by,

$$\mathbf{y} = \mathbf{H}\mathbf{x}_{\mathrm{CF}} + \boldsymbol{\eta} \tag{3.53}$$

$$= \mathbf{HAWs} + \mathbf{H}\boldsymbol{\varepsilon} + \boldsymbol{\eta}, \tag{3.54}$$

(where the subcarrier and iteration indices are dropped for clarity), whilst at a given user the received signal is

$$y_k = \mathbf{h}_k^T \mathbf{A}\mathbf{w}_k s_k + \sum_{j \neq k} \mathbf{h}_k^T \mathbf{A}\mathbf{w}_j s_j + \mathbf{h}_k^T \boldsymbol{\varepsilon} + \eta. \tag{3.55}$$

The clipping manifests itself in two effects:

- The Bussgang attenuation, $\mathbf{A}$, distorts the MIMO precoding. This will generally lead to a reduction in received signal power for the intended user data stream, and an increase in interference from other user data streams. For example, under zero-forcing precoding the orthogonality of the data streams will be lost if $\mathbf{A}$ is not of the form $\alpha \mathbf{I}_M$.
- The clipping noise introduces an additional source of (approximately) Gaussian error at the receiver, with power $\mathbf{h}_k^T \mathbf{R}_\varepsilon \mathbf{h}_k^*$. For a given $\mathbf{R}_\varepsilon$, the strength of the received clipping noise depends on both the 'direction' of the user channel, $\mathbf{h}_k / \|\mathbf{h}_k\|$, and the strength of the channel vector, $\|\mathbf{h}_k\|$.

Using the fact that the signal components are all uncorrelated, the receive SINR is

$$\mathrm{SINR}_k = \frac{\rho_k \left|\mathbf{h}_k^T \mathbf{A}\mathbf{w}_k\right|^2}{\sum_{j\neq k} \rho_j \left|\mathbf{h}_k^T \mathbf{A}\mathbf{w}_j\right|^2 + \mathbf{h}_k^T \mathbf{R}_\varepsilon \mathbf{h}_k^* + 1}, \tag{3.56}$$

and

$$\mathcal{C}_k = \log_2\left(1 + \mathrm{SINR}_k\right) \tag{3.57}$$

is an achievable rate (since treating $\mathbf{h}_k^T \boldsymbol{\varepsilon}$ as Gaussian noise provides a lower bound on capacity [84]).

Since clipping reduces the average power of the transmit signal, for a meaningful performance comparison the clipped and filtered signal should be normalised so that it has the same average power as the original OFDM signal. Figure 3.9 compares the mean user capacity with and without average power normalisation at different clipping ratios, for a system with 8 users, where the user channels have identical pathloss and ten Rayleigh fading channel taps. The power level is set so that each user has an SNR of 15 dB without clipping. The clipping error covariance matrices for each subcarrier are calculated numerically. In this example, most of the capacity loss is due to the Bussgang gain reducing the desired signal power, and is partially compensated for by normalising the transmit power.

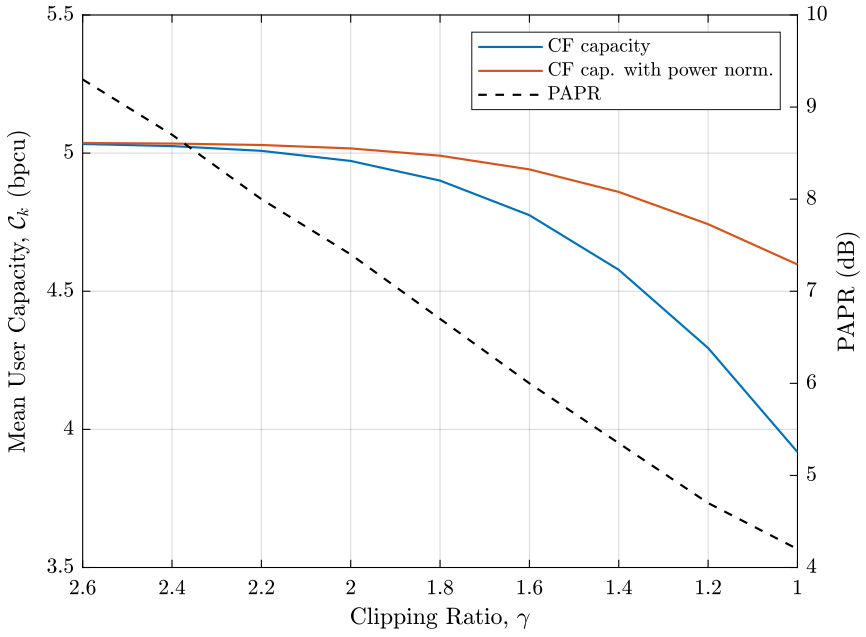

Figure 3.9: Mean user capacity for different clipping ratios with and without power normalisation, clipping & filtering with three iterations. Normalised to receive SNR 15 dB (unclipped), 8 users, 64 BS antennas, 512 subcarriers.

3.4.1 Asymptotic Performance

As discussed in Chapter 2, much MIMO analysis assumes fading at different antennas occurs independently. Under this assumption, providing significant frequency diversity is present[3], it can then be assumed that transmit power is evenly distributed between antennas, with low correlations between their time domain transmit samples,

$$\mathbf{R}_x \approx \frac{P_T}{M}\mathbf{I}_M. \tag{3.58}$$

The clipped & filtered signals will then also be uncorrelated, and therefore

$$\mathbf{A} = \alpha\mathbf{I}_M, \tag{3.59}$$

with spatially white clipping noise

$$\mathbf{R}_\varepsilon = \sigma_\varepsilon^2\mathbf{I}_M. \tag{3.60}$$

For large M, the received clipping noise power is given by,

$$\mathbf{h}_k^T\mathbf{R}_\varepsilon\mathbf{h}_k^* \approx \sigma_\varepsilon^2\beta_k M. \tag{3.61}$$

For a constant clipping ratio the clipping noise power is proportional to the transmit power, $\sigma_\varepsilon^2 \propto P_T/M$. Since, due to array gain, the required transmit power also decreases $P_T \sim 1/M$ (cf. Section 2.5.1), the received clipping noise asymptotically disappears as $M \to \infty$,

$$\mathbf{h}_k^T\mathbf{R}_\varepsilon\mathbf{h}_k^* \to 0. \tag{3.62}$$

The effect of clipping is then only to reduce the SNR at the users by a factor α, which can be compensated for by normalising the final signal power. This is illustrated in Figure 3.10 for 8 users with independent 10-tap Rayleigh fading channels and 3 iterations of clipping & filtering with $\gamma = 1.2$. Observe here that the impact of clipping on performance is more significant at higher SNRs, with receiver noise being the main limiting factor at low SNR.

[3]If the channels to all antennas have the same path loss, and the precoding matrices are independent on different blocks of subcarriers then by the law of large numbers $\frac{1}{N}\sum_{l=0}^{N-1}\mathbf{W}_l\mathbf{W}_l^\dagger \to \frac{P_T}{M}\mathbf{I}_M$ for large N.

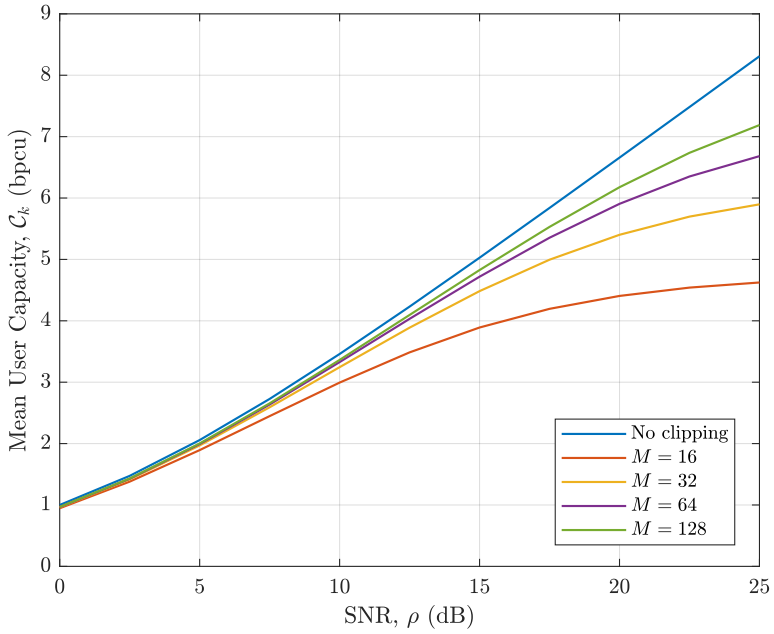

Figure 3.10: Mean user capacity with varying antenna numbers, 8 users, 512 subcarriers, $\gamma = 1.2$, 3 iterations.

3.4.2 Spatially Correlated Channels

As shown in Figure 3.6, precoding for spatially correlated MIMO channels can result in increased PAPR, and a greater need for PAPR reduction. This change in channel statistics will also impact the effect clipping & filtering has on performance.

Firstly, when there is variation in per-antenna average transmit power, P_m, the Bussgang gain matrix, $\mathbf{A}$, will not be of the form $\alpha \mathbf{I}_M$, and the symbol precoding will be distorted – leading to additional inter-user interference.

Secondly, high spatial correlation in the OFDM transmit signal samples will lead to spatially correlated clipping noise. The relationship between the OFDM signal spatial correlation and clipping noise spatial correlation is difficult to analyse, and the clipping noise correlation must generally be evaluated numerically

$$\mathbf{R}_{\varepsilon,l} = \mathbb{E}\big[\big(\mathbf{x}_{\mathrm{CF},l} - \mathbf{A}\mathbf{x}_l\big)\big(\mathbf{x}_{\mathrm{CF},l} - \mathbf{A}\mathbf{x}_l\big)^{\dagger}\big]. \tag{3.63}$$

However, analysis of the spatial characteristics of general non-linear distortion in [147] shows that when there is a dominant transmit direction the distortion will tend to be beamformed in that direction, whilst for multiple transmit directions, with none dominant, the distortion will tend to be radiated fairly isotropically (i.e. low spatial correlation). It also shown that the number of directions the distortion is radiated in increases rapidly with the number of users ($\sim K^3$), whilst frequency diversity will tend to reduce the distortion correlation.

The spatial correlation of a signal can be usefully visualised using the eigenvalues of the spatial covariance matrix [15]. The magnitudes of the eigenvalues of both the transmit signal correlation matrix, $\mathbf{R}_x$, and the clipping noise covariance matrix, $\mathbf{R}_{\varepsilon,l}$, are shown in Figure 3.11

(all normalised to their largest respective eigenvalue). Two channels are considered – the i.i.d Rayleigh fading channel with 10 taps, and the correlated Rayleigh fading channel from Figure 3.6. Conventional ZF precoding with a total power constraint is used, and the eigenvalues are averaged over many fading realisations.

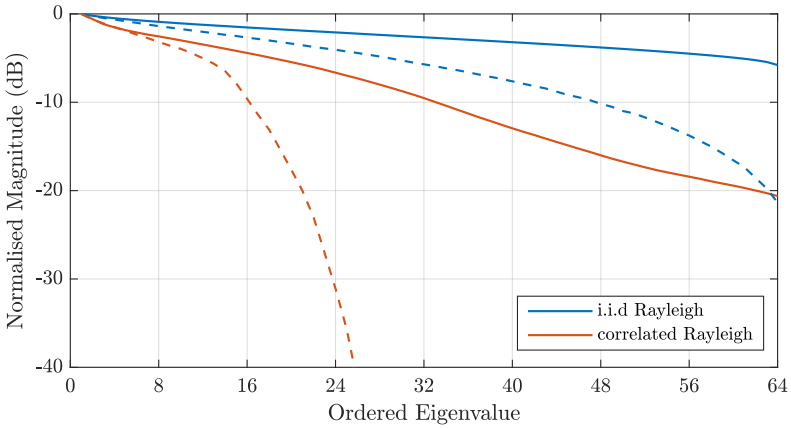

Figure 3.11: Solid line: normalised eigenvalues of clipping noise covariance (measured on central subcarrier), dashed line: normalised eigenvalues of transmit signal covariance.

The eigenvalue spread for the clipping noise covariance under i.i.d fading is small, indicating that it is radiated almost isotropically. Under correlated fading, the transmit signal samples are highly correlated. The clipping noise is radiated more evenly than the correlated transmit signal, but exhibits significantly higher spatial correlation than the clipping noise produced under i.i.d fading. Overall, the impact of clipping on performance can be expected to be more severe in the correlated fading case, as is shown in Figure 3.12.

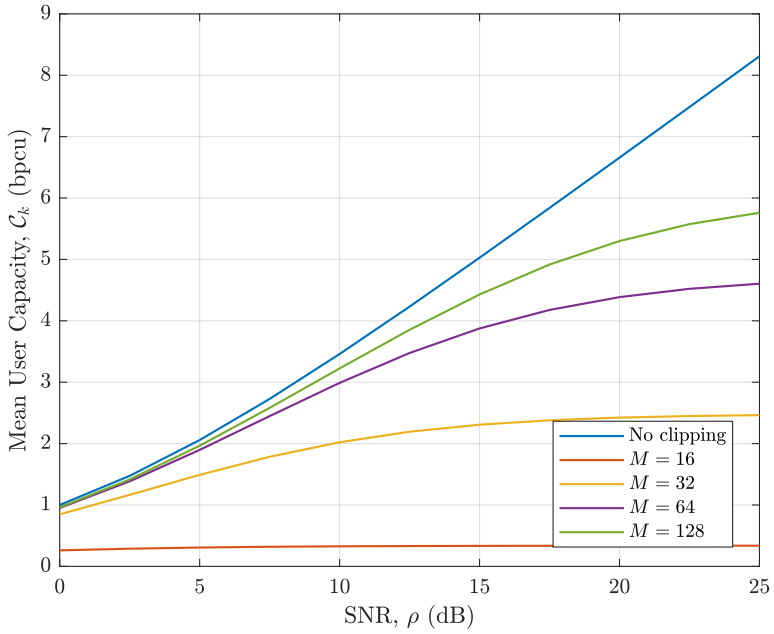

Figure 3.12: Mean user capacity with varying antenna numbers, 8 users, 512 subcarriers, $\gamma = 1.2$, 3 iterations.

3.4.3 Near-Far Effect

The power of the received clipping noise at user k is proportional to the channel strength, $\|\mathbf{h}_k\|^2$. Under max-min power control a near-far effect can therefore occur where, in the presence of users with weak channels, the clipping noise strongly degrades the performance of users with strong channels. This occurs because under max-min power control, more power is allocated to users with weak channels, so that all users have the same received signal strength. The clipping noise (which has a power related to the total transmit power) therefore has a disproportionate effect on the near users. This is illustrated in Figure 3.13, where the near user has a channel strength 10 dB greater than the far users – the degradation of the performance of the near user due to clipping noise is much worse than the far users[4].

When all users have a similar pathloss this effect is reduced, because the BS power is shared more equally between users. For example if all the users have strong channels, under max-min power control they will each receive a stronger signal (compared to the case where far users are also present), and the impact of the clipping noise will be reduced. The spatial filtering method proposed in the next section presents an attractive means of addressing this problem, whilst also improving the performance of the far users.

[4] A similar near-far effect occurs when beamforming with imperfect CSI, and can be addressed by incorporating CSI errors into max-min power control. A similar thing could be potentially attempted with clipping, and is a possible area for future work.

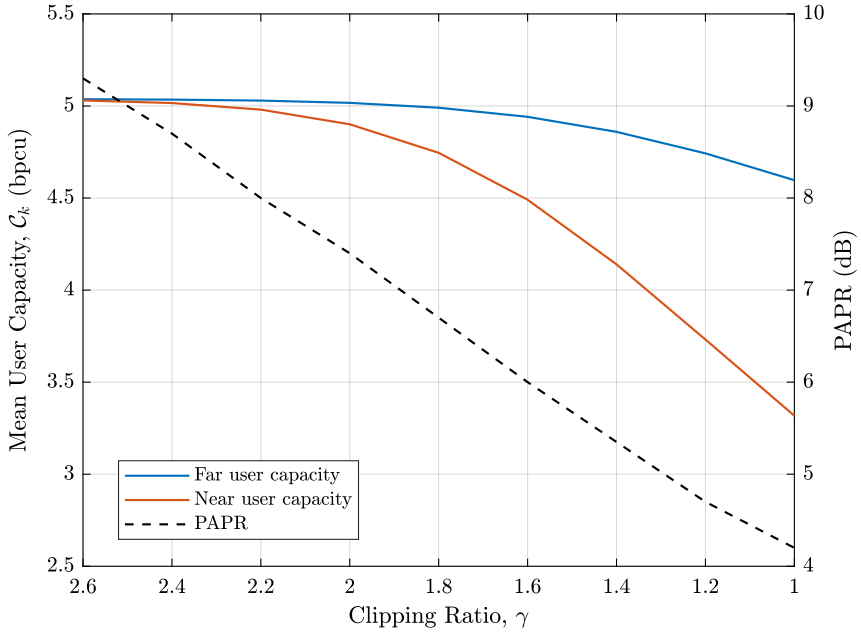

Figure 3.13: Near-far effect of clipping on user capacity where average channel strength of near user is 10 dB greater than far users. Power normalised to receive SNR 15 dB (unclipped), three iterations of clipping & filtering, 1 near user, 7 far users, 64 BS antennas, 512 subcarriers.

3.5 Clipping & Least Squares Spatial Filtering

The use of aggressive clipping & filtering leads to transmit signals with very good PAPR performance, at the expense of often considerable degradation to link performance. If a signal can be found that closely represents the CF signal in the time domain, but which gives improved link performance, this signal will be a strong candidate for a low PAPR transmit signal. This section shows that such a signal can be found through spatial filtering of the CF signal.

The least squares spatial filtering approach is based on the idea, previously noted in [219], that a signal that closely matches the CF signal in a squared error sense, i.e. a transmit signal, $\mathbf{x}_{\text{LS}}[n]$ with a low value of

$$\epsilon = \frac{1}{qN} \sum_{n=0}^{qN-1} \left\| \mathbf{x}_{\text{LS}}[n] - \mathbf{x}_{\text{CF}}[n] \right\|^2, \tag{3.64}$$

will also display attractive PAPR characteristics. The use of a squared error measure is attractive since, as this section shows, it enables a simple linear closed form expression for the transmit signal to be found. Whilst its use is heuristic – it does not explicitly minimise the peak power of the transmit signal – the results presented here show that the least squares filtering approach can produce signals with very low PAPR, whilst suffering only a small performance loss.

This section begins by outlining the conventional spatial filtering, or nullspace projection, method that has been previously been employed in [205], [136] & [219]. Using the Bussgang clipping model developed in the previous section, the limitations of this scheme are identified. A new scheme that accounts properly for the effects of clipping – referred to as the Bussgang-aware

least squares (BLS) method – is then developed, and the benefits of this scheme demonstrated. The idea of active constellation extension is then incorporated into this scheme.

3.5.1 Least Squares Spatial Filtering / Nullspace Projection Method

A good candidate transmit signal achieves good link performance whilst closely approximating the low PAPR CF signal. A key feature of the squared error measure (3.64) is that, using Parseval's theorem, it may also be expressed in terms of the squared error on the subcarriers,

$$\epsilon = \frac{1}{N}\sum_{l=0}^{N-1} \left\|\mathbf{x}_{\mathrm{LS},l} - \mathbf{x}_{\mathrm{CF},l}\right\|^2. \tag{3.65}$$

Minimisation of the squared error between the transmit and CF time domain signals can therefore be achieved by minimising the error on each subcarrier – which is attractive since MIMO-OFDM processing is carried out on a subcarrier level. In the previously proposed methods, the transmit signal is chosen such that the signal received at the users is equal to the original precoded signal, i.e.

$$\mathbf{H}_l\mathbf{x}_{\mathrm{LS},l} = \mathbf{H}_l\mathbf{x}_l. \tag{3.66}$$

The optimal transmit signal can be found independently for each subcarrier, by solving N parallel constrained least squares optimisations

$$\underset{\mathbf{x}_{\mathrm{LS},l}}{\text{minimise}} \quad \left\|\mathbf{x}_{\mathrm{LS},l} - \mathbf{x}_{\mathrm{CF},l}\right\|^2, \tag{3.67}$$

$$\text{subject to} \quad \mathbf{H}_l\mathbf{x}_{\mathrm{LS},l} = \mathbf{H}_l\mathbf{x}_l. \tag{3.68}$$

Dropping the subcarrier index for clarity, the solution to this can be easily shown to be (see Appendix 1.1)

$$\mathbf{x}_{\mathrm{LS}} = \mathbf{x}_{\mathrm{CF}} - \mathbf{H}^\dagger(\mathbf{H}\mathbf{H}^\dagger)^{-1}\mathbf{H}\big(\mathbf{x}_{\mathrm{CF}} - \mathbf{x}\big). \tag{3.69}$$

The CF signal can be written as the sum of the OFDM signal, $\mathbf{x}$, and an additive error as in [205], [136], [219],

$$\mathbf{x}_{\mathrm{CF}} = \mathbf{x} + \mathbf{c}, \tag{3.70}$$

where the additive error, $\mathbf{c}$, is responsible for cancelling the peaks in the OFDM signal. The least squares solution can then be interpreted intuitively in two ways.

First, the transmit signal can be written

$$\mathbf{x}_{\mathrm{LS}} = \mathbf{x} + \big(\mathbf{I}_M - \mathbf{H}^\dagger(\mathbf{H}\mathbf{H}^\dagger)^{-1}\mathbf{H}\big)\mathbf{c}. \tag{3.71}$$

The matrix $\big(\mathbf{I}_M - \mathbf{H}^\dagger(\mathbf{H}\mathbf{H}^\dagger)^{-1}\mathbf{H}\big)$ is the orthogonal nullspace projection matrix for $\mathbf{H}$, and the least squares approximation is the original OFDM signal plus a peak reduction signal generated through spatially filtering the CF peak cancellation signal (by projecting it into the channel nullspace). This is the approach used in [205], [136] & [219], where the nullspace projection matrix is calculated using the singular value decomposition of $\mathbf{H}$.

Alternatively, the transmit signal can be expressed

$$\mathbf{x}_{\mathrm{LS}} = \mathbf{x}_{\mathrm{CF}} - \mathbf{W}^{(\mathrm{ZF})}\mathbf{d} \tag{3.72}$$

where $\mathbf{W}^{(\mathrm{ZF})}$ is the conventional zero-forcing precoder as described in Section 2.3, and $\mathbf{d}$ is the error at the receivers due to the peak reduction signal,

$$\mathbf{d} = \mathbf{H}\mathbf{c}. \tag{3.73}$$

The least squares signal can then be thought of as the CF signal plus a second signal that cancels out the error caused by the peak reduction signal. This second interpretation is useful from an implementation perspective, showing that the least squares signal can be generated using the ZF precoder – which in many cases will already be available.

Analysis

The addition of the error cancellation signal to the original CF signal changes the transmit signal samples, with squared error

$$\epsilon = \frac{1}{N}\sum_{l=0}^{N-1} \left\|\mathbf{W}_l^{(\mathrm{ZF})}\mathbf{d}_l\right\|^2. \tag{3.74}$$

When this error is small, the transmit signal can be expected to retain its good PAPR properties, and when it is large peak regrowth can be expected.

A useful insight is provided by analysing the expected transmit signal error

$$\begin{aligned}
\mathbb{E}[\epsilon] &= \frac{1}{N}\sum_{l=0}^{N-1} \mathbb{E}\left[\left\|\mathbf{W}_l^{(\mathrm{ZF})}\mathbf{d}_l\right\|^2\right] && (3.75)\\
&= \frac{1}{N}\sum_{l=0}^{N-1} \mathbb{E}\left[\left\|\mathbf{W}_l^{(\mathrm{ZF})}\mathbf{H}_l\left(\boldsymbol{\varepsilon}_l + \left(\mathbf{A}-\mathbf{I}_M\right)\mathbf{W}_l\mathbf{s}_l\right)\right\|^2\right] && (3.76)\\
&= \frac{1}{N}\sum_{l=0}^{N-1} \left\|\mathbf{W}_l^{(\mathrm{ZF})}\mathbf{H}_l\mathbf{R}_{\varepsilon,l}^{1/2}\right\|^2 + \frac{1}{N}\sum_{l=0}^{N-1} \rho\left\|\mathbf{W}_l^{(\mathrm{ZF})}\mathbf{H}_l\left(\mathbf{A}-\mathbf{I}_M\right)\mathbf{W}_l\right\|^2 && (3.77)\\
&= \frac{1}{N}\sum_{l=0}^{N-1} \left\|\mathbf{W}_l^{(\mathrm{ZF})}\mathbf{H}_l\mathbf{R}_{\varepsilon,l}^{1/2}\right\|^2 + \frac{1}{N}\sum_{l=0}^{N-1} \rho\left\|\mathbf{A}\mathbf{W}_l - \mathbf{W}_l\right\|^2 && (3.78)
\end{aligned}$$

where the second line follows since $\mathbf{c}_l = \mathbf{A}\mathbf{x}_l + \boldsymbol{\varepsilon}_l - \mathbf{x}_l$, and the fourth line since $\mathbf{W}_l^{(\mathrm{ZF})}\mathbf{H}_l$ is an orthogonal projection into the channel column space (which the precoded signal component lies fully within).

This error is a combination of the power needed to cancel the uncorrelated clipping noise and the power needed to correct the error in the data symbol precoding due to the Bussgang gain. It is the power required to correct the symbol precoding that limits the PAPR performance of this scheme. This can be understood by considering the simplified example from 3.4.1, with

$\mathbf{A} = \alpha \mathbf{I}_M$. Using (3.14), the power required to cancel the precoding error component is

$$\frac{1}{N} \sum_{l=0}^{N-1} \rho \left\| (\alpha - 1) \mathbf{W}_l \right\|^2 = (1 - \alpha)^2 P_T, \tag{3.79}$$

where P_T is the total transmit power of the original OFDM signal as in (3.15). As clipping is applied more aggressively (γ decreases), the Bussgang gain, α, decreases and ϵ increases. When α is small the power of the cancellation signal is comparable to the power of the original OFDM signal, and significant peak regrowth occurs – effectively reversing the PAPR reduction achieved by clipping.

This is shown in Figure 3.14, where beyond a clipping ratio of around $\gamma = 1.5$ significant peak regrowth occurs and the PAPR grows again. This limits the PAPR reduction that can be achieved, even when many clipping & spatial filtering iterations are used.

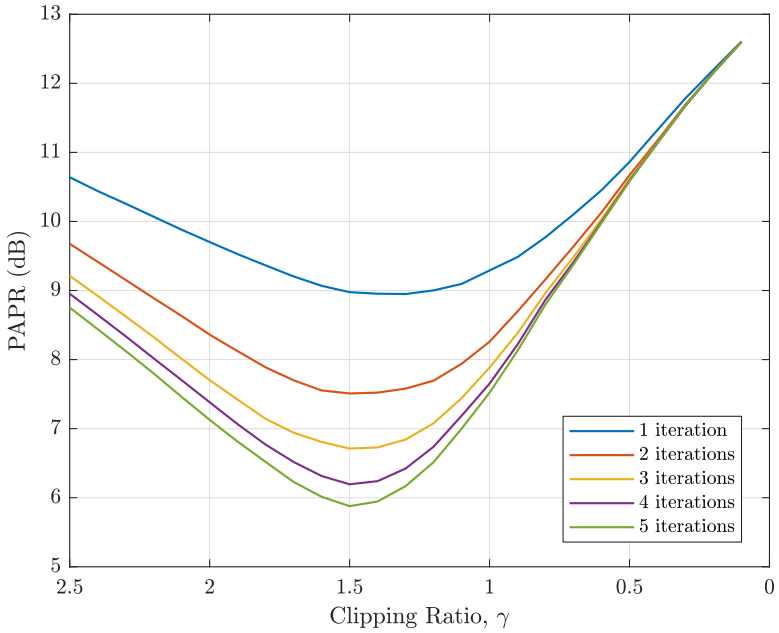

Figure 3.14: PAPR reduction of iterative least squares filtering scheme in Rayleigh channel, $K = 8$, $M = 64$ and $N = 512$.

3.5.2 Bussgang-aware Least Squares Filtering

The conventional least squares filtering method attempts to maintain the same received signal at the users, despite the fact that – as shown in Section 3.3 – the clipping operation inherently reduces the strength of the OFDM signal component. This prevents the method from converging when the clipping ratio is small, limiting the PAPR reduction it can achieve. Here, it is proposed that this can be addressed by choosing a target received signal that accounts for the loss in gain – the Bussgang-aware least squares filtering approach (BLS).

The key idea with the Bussgang-aware least squares approach is to replace the constraints

in (3.66) with

$$\mathbf{H}_l\mathbf{x}_{\mathrm{BLS},l} = \mu_l\mathbf{H}_l\mathbf{x}_l \tag{3.80}$$

where μ_l are constants chosen to account for the loss in gain due to clipping. The least squares signals are then the solutions to

$$\underset{\mathbf{x}_{\mathrm{BLS},l}}{\text{minimise}} \quad \left\|\mathbf{x}_{\mathrm{BLS},l} - \mathbf{x}_{\mathrm{CF},l}\right\|^2, \tag{3.81}$$

$$\text{subject to} \quad \mathbf{H}_l\mathbf{x}_{\mathrm{BLS},l} = \mu_l\mathbf{H}_l\mathbf{x}_l, \tag{3.82}$$

which are given, similarly, as

$$\mathbf{x}_{\mathrm{BLS}} = \mathbf{x}_{\mathrm{CF}} - \mathbf{W}^{(\mathrm{ZF})}\mathbf{H}\big(\mathbf{x}_{\mathrm{CF}} - \mu\mathbf{x}\big) \tag{3.83}$$

$$= \mathbf{x}_{\mathrm{CF}} - \mathbf{W}^{(\mathrm{ZF})}\mathbf{H}\big(\mathbf{AWs} + \boldsymbol{\varepsilon} - \mu\mathbf{Ws}\big). \tag{3.84}$$

The cancellation signal is now the difference between the clipped signal and a *scaled* version of the original OFDM signal.

The optimal choices of scaling factor, in terms of PAPR, are the ones that minimise the average power of the cancellation signal,

$$\underset{\mu_l \forall l}{\text{minimise}} \quad \mathbb{E}[\epsilon], \tag{3.85}$$

equivalent to solving the N optimisation problems

$$\underset{\mu_l}{\text{minimise}} \quad \left\|\mathbf{AW}_l - \mu_l\mathbf{W}_l\right\|^2. \tag{3.86}$$

This is achieved by the scaling factors (see Appendix 1.2)

$$\mu_l = \frac{\mathrm{Tr}\left(\mathbf{W}_l^\dagger\mathbf{AW}_l\right)}{\mathrm{Tr}\left(\mathbf{W}_l^\dagger\mathbf{W}_l\right)} \tag{3.87}$$

The cancellation signal now has two roles – it removes the clipping noise at the users, and it corrects the distortion to the precoding caused by clipping, but it does not attempt to compensate for the attenuation of the OFDM signal component. As a result, a better least squares approximation is achieved than the conventional least squares method, and significantly less peak regrowth occurs, as shown for the 10 tap Rayleigh channel in Figure 3.15.

For the illustrative case of $\mathbf{A} = \alpha\mathbf{I}_M$, no power is expended correcting the precoding under the BLS scheme (since $\mu = \alpha$) whilst for $\mathbf{R}_\varepsilon = \sigma_\varepsilon^2\mathbf{I}_M$, the power required to cancel the clipping noise is given by,

$$\frac{1}{N}\sum_{l=0}^{N-1}\left\|\mathbf{W}_l^{(\mathrm{ZF})}\mathbf{H}_l\mathbf{R}_{\varepsilon,l}^{1/2}\right\|^2 = K\sigma_\varepsilon^2, \tag{3.88}$$

and by the reasoning used in Section 3.4.1, under this model as $M \to \infty$,

$$K\sigma_\varepsilon^2 \to 0. \tag{3.89}$$

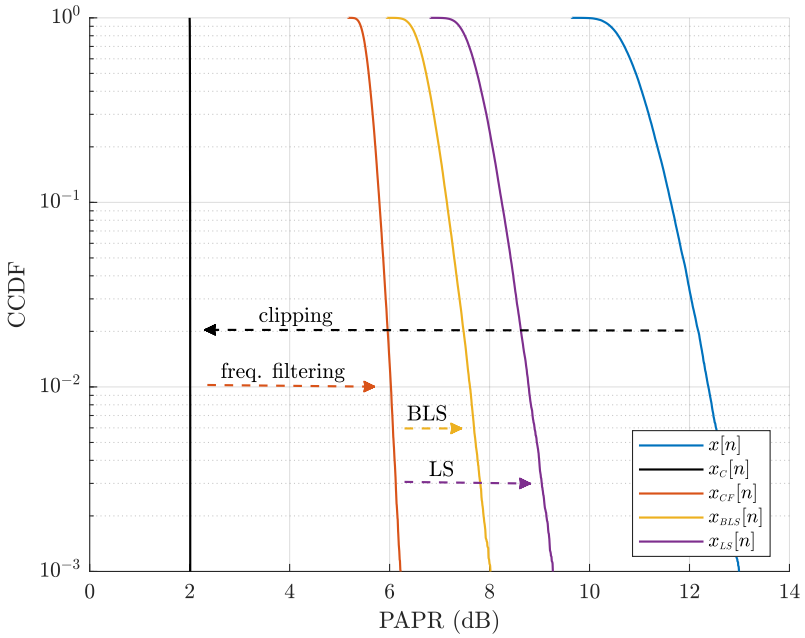

Figure 3.15: Comparison of PAPR regrowth of LS and BLS spatial filtering methods.

Thus, if sufficiently rich scattering is present, as the number of antennas grows it can be expected that the least squares approximation of the clipped & filtered signal becomes more accurate, and the PAPR performance improves.

3.5.3 Iterative BLS

To achieve good performance the iterative clipping & filtering should be used, with the BLS approximation applied after each stage. Using the generalised Bussgang model, the clipped signal on subsequent iterations can be written

$$\begin{aligned}\mathbf{x}_{\mathrm{CF}}^{(i)} &= \mathbf{A}^{(i)}\mathbf{x}_{\mathrm{BLS}}^{(i-1)} + \boldsymbol{\varepsilon}^{(i)} && (3.90)\\ &= \mu^{(i-1)}\mathbf{A}^{(i)}\mathbf{x} + \mathbf{A}^{(i)}\big(\mathbf{I}_M - \mathbf{W}^{(\mathrm{ZF})}\mathbf{H}\big)\boldsymbol{\varepsilon}^{(i-1)} + \boldsymbol{\varepsilon}^{(i)} && (3.91)\\ &= \mu^{(i-1)}\mathbf{A}^{(i)}\mathbf{x} + \boldsymbol{\varepsilon}', && (3.92)\end{aligned}$$

where $\mathbf{A}^{(i)}$ is given by

$$\mathbf{A}^{(i)} = \mathbf{R}_{x_{cf},x_{bls}}^{(i)}\left(\mathbf{R}_{x_{bls}}^{(i-1)}\right)^{-1}, \tag{3.93}$$

with

$$\mathbf{R}_{x_{cf},x_{bls}}^{(i)} = \mathbb{E}\big[\mathbf{x}_{\mathrm{CF}}^{(i)}[n]\mathbf{x}_{\mathrm{BLS}}^{(i-1)\dagger}[n]\big], \tag{3.94}$$

$$\mathbf{R}_{x_{bls}}^{(i-1)} = \mathbb{E}\big[\mathbf{x}_{\mathrm{BLS}}^{(i-1)}[n]\mathbf{x}_{\mathrm{BLS}}^{(i-1)\dagger}[n]\big]. \tag{3.95}$$

Estimating $\mathbf{A}^{(i)}$ fully is prohibitive, since it requires the estimation and inversion of an $M \times M$ matrix. Instead, the Bussgang gain may be approximated as a diagonal matrix, $\tilde{\mathbf{A}}^{(i)}$, similarly

to (3.30), with diagonal entries

$$\left[\tilde{\mathbf{A}}^{(i)}\right]_{m,m} = \frac{\mathbb{E}\left[x_{\mathrm{CF},m}^{(i)}[n]x_{\mathrm{BLS},m}^{(i-1)*}[n]\right]}{\mathbb{E}\left[\left|x_{\mathrm{BLS},m}^{(i-1)}[n]\right|^2\right]} \tag{3.96}$$

$$\approx \frac{\sum_{n=0}^{qN-1} x_{\mathrm{CF},m}^{(i)}[n]x_{\mathrm{BLS},m}^{(i-1)*}[n]}{\sum_{n=0}^{qN-1}\left|x_{\mathrm{BLS},m}^{(i-1)}[n]\right|^2}. \tag{3.97}$$

Since the Bussgang matrix estimate is only used for setting the target symbol scaling some error can be tolerated provided the overall effect of signal attenuation is still captured.

The Bussgang least squares filtering on the second iteration is then applied as,

$$\mathbf{x}_{\mathrm{BLS}}^{(i)} = \mathbf{x}_{\mathrm{CF}}^{(i)} - \mathbf{W}^{(\mathrm{ZF})}\mathbf{H}\left(\mathbf{x}_{\mathrm{CF}}^{(i)} - \mu^{(i)}\mathbf{x}\right) \tag{3.98}$$

where

$$\mu^{(i)} = \frac{\mathrm{Tr}\left(\mathbf{W}^{\dagger}\tilde{\mathbf{A}}^{(i)}\mathbf{W}\right)}{\mathrm{Tr}\left(\mathbf{W}^{\dagger}\mathbf{W}\right)}\mu^{(i-1)} \tag{3.99}$$

Figure 3.16 shows that, unlike the conventional least squares method, with this strategy the scheme converges at all clipping ratios. Furthermore, significantly lower PAPR can be achieved – a PAPR of less than 5 dB can be achieved, compared to around 6 dB for the least squares method.

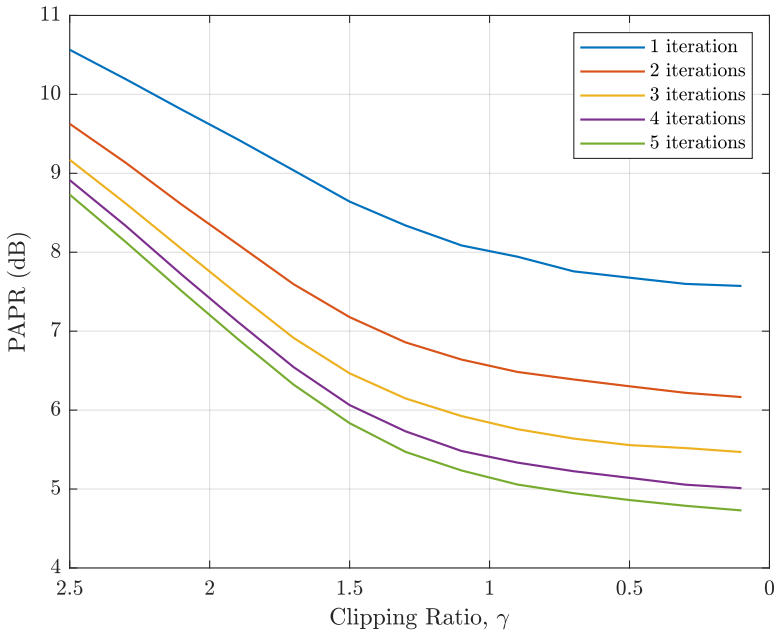

Figure 3.16: PAPR reduction of iterative BLS scheme in Rayleigh channel, $K = 8$, $M = 64$ and $N = 512$.

An apparent downside of the BLS PAPR reduction scheme is that it reduces the strength of the signal received at the users. However, the BLS scheme also reduces the average power of the transmit signal, and after power normalisation the loss in overall receiver SNR is small, as

shown in Figure 3.17.

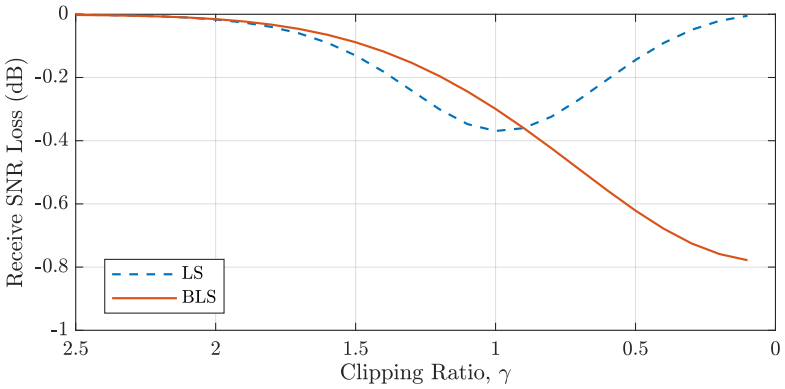

Figure 3.17: Loss in received signal power after power normaliation. 5 clipping iterations, i.i.d Rayleigh channel, $K = 8$, $M = 64$ and $N = 512$.

At a clipping ratio $\gamma = 1$, the BLS scheme only imposes a 0.3 dB penalty relative to unclipped OFDM. Comparing with Figure 3.16, beyond this point the PAPR continues to reduce, but slowly, whilst the SNR loss drops off more quickly, indicating that operating $\gamma = 1$ is a good operating point. Note that the LS method does not achieve PAPR reduction at small clipping ratios. Figure 3.18 shows the PAPR convergence and SNR loss of the BLS method. The use of phase-only clipping results in significant SNR loss compared to standard clipping.

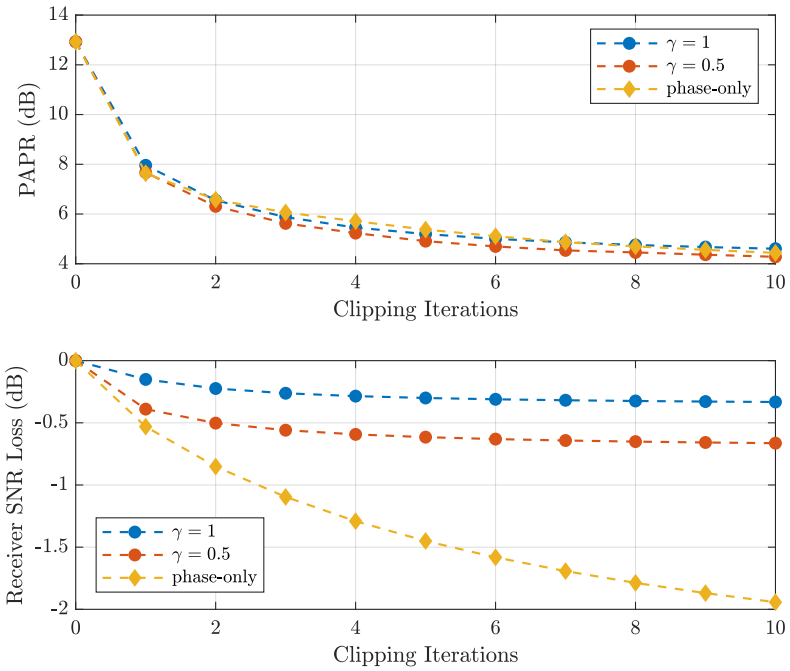

Figure 3.18: PAPR reduction and receive SNR loss of iterative BLS clipping & filtering in 10 tap i.i.d Rayleigh channel, $K = 8$, $M = 64$ and $N = 512$.

Since the projection matrix $\mathbf{W}_l^{(\mathrm{ZF})}\mathbf{H}_l$ does *not* depend on the user channel strengths[5] the

[5]This can be seen by writing $\mathbf{H} = \mathbf{D}\bar{\mathbf{H}}$ where $\mathbf{D} = \mathrm{diag}(\|\mathbf{h}_k\|)$ and the rows of $\bar{\mathbf{H}}$ have unit norm. Then

near-far clipping noise effects discussed in 3.4.3 are eliminated. Figure 3.19 shows the BER for uncoded QPSK in the near-far scenario where one user has a average channel 10 dB stronger than the others. Under the BLS scheme the performance loss is negligible, but under conventional clipping & filtering the near user experiences a high error floor.

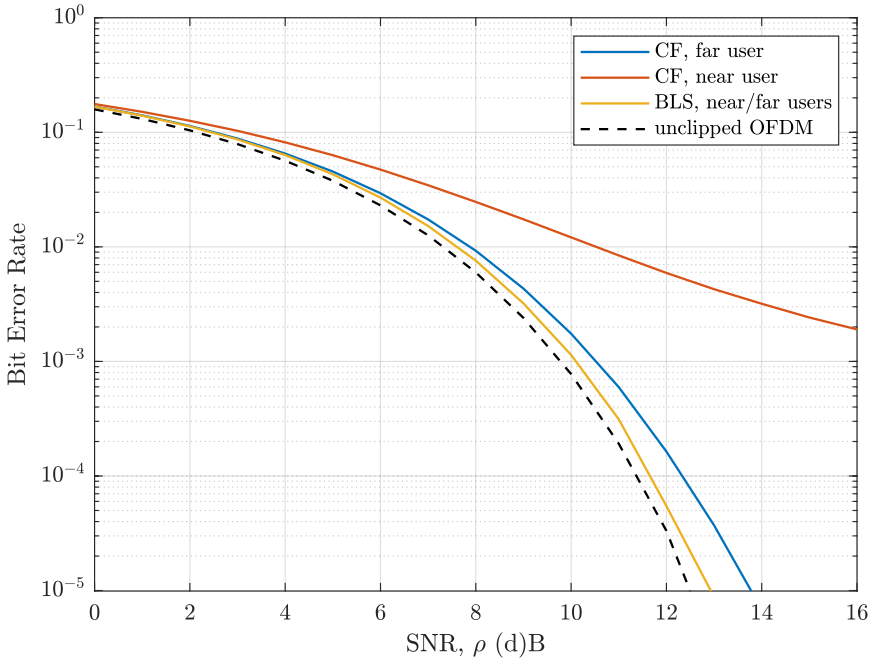

Figure 3.19: Bit error rate in near-far scenario, i.i.d Rayleigh channel, QPSK, 5 clipping iterations with $\gamma = 1$, $K = 8$, $M = 64$ and $N = 512$.

3.5.4 BLS in Correlated Channels

The error in the least squares approximation depends on the Bussgang gain and clipping noise covariance, and therefore the PAPR reduction that is achieved depends on the MIMO channel. When operating in spatially correlated channels, clipping & filtering can introduce more inter-user interference and clipping noise at the users, as discussed in Section 3.4.2. Under the BLS scheme, this means that more power must be used to cancel the effects of clipping, which will generally result in higher overall PAPR.

Figure 3.20 shows the PAPR reduction and SNR loss (compared to unclipped OFDM) achieved when iteratively applying the BLS clipping & filtering scheme, for the correlated channel from Section 3.4.2. With clipping ratio $\gamma = 1$ the scheme can reduce the PAPR from over 15 dB to less than 6 dB with less than 0.4 dB performance penalty, but requires a larger number of iterations than in the i.i.d Rayleigh channel.

$\mathbf{W}^{(\mathrm{ZF})}\mathbf{H} = \bar{\mathbf{H}}^{\dagger}\left(\bar{\mathbf{H}}\bar{\mathbf{H}}^{\dagger}\right)^{-1}\bar{\mathbf{H}}$ – any scaling of $\|\mathbf{h}_k\|$ does not impact the power required to cancel the user clipping noise.

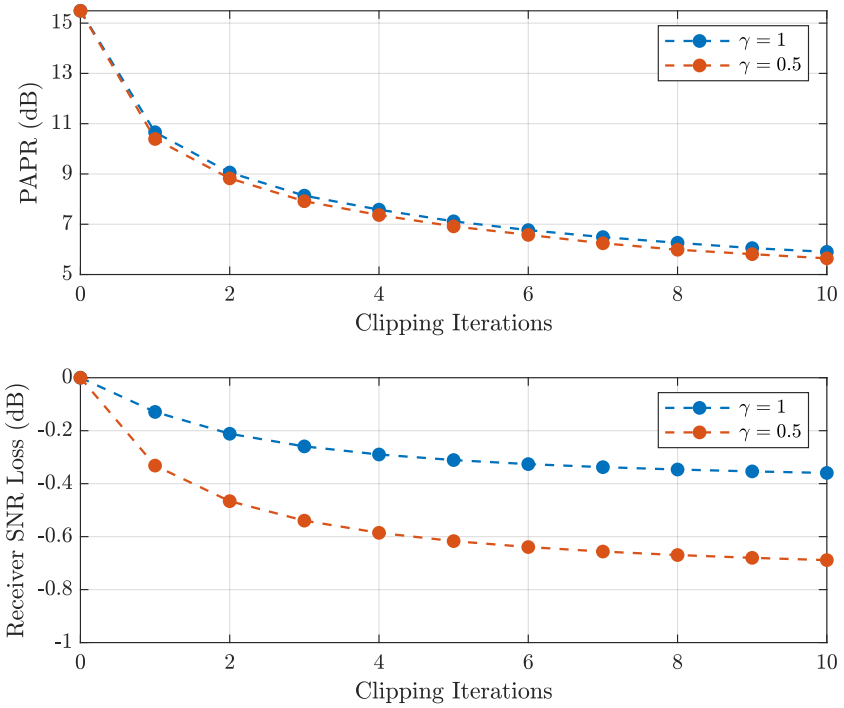

Figure 3.20: PAPR reduction and receive SNR loss of iterative BLS clipping & filtering in correlated Rayleigh channel, $K = 8$, $M = 64$ and $N = 512$.

3.5.5 Active Constellation Extension

Under ACE some clipping noise is permitted at the receiver providing it falls within certain regions of the symbol constellation [112]. This enables a closer representation of the CF signal to be used, improving PAPR performance.

The BLS is easily extended to include ACE by replacing the constraint in (3.80) with

$$\mathbf{H}\mathbf{x}_{\text{BLS}} = \mu \mathbf{H}\mathbf{W}\tilde{\mathbf{s}}, \tag{3.100}$$

where $\tilde{\mathbf{s}}$ is the nearest point to the clipped signal that lies within the extended constellation region,

$$\underset{\tilde{\mathbf{s}}}{\text{minimise}} \quad \left\|\mathbf{H}\mathbf{x}_{\text{CF}} - \mu \mathbf{H}\mathbf{W}\tilde{\mathbf{s}}\right\|^2, \tag{3.101}$$

$$\text{subject to} \quad \tilde{\mathbf{s}} \in C(\mathbf{s}), \tag{3.102}$$

where $C(\mathbf{s})$ is the extended constellation region for symbol vector $\mathbf{s}$. This is achieved by projecting the received clipped signal point for each user into the permitted constellation region (with appropriate scaling), as illustrated geometrically for QPSK signalling in Figure 3.21.

The extended constellation point, $\tilde{\mathbf{s}}$ may contain either/both clipping noise and inter-user interference, and necessarily have

$$\left\|\mathbf{H}\mathbf{x}_{\text{CF}} - \mu \mathbf{H}\mathbf{W}\tilde{\mathbf{s}}\right\|^2 \leq \left\|\mathbf{H}\mathbf{x}_{\text{CF}} - \mu \mathbf{H}\mathbf{W}\mathbf{s}\right\|^2, \tag{3.103}$$

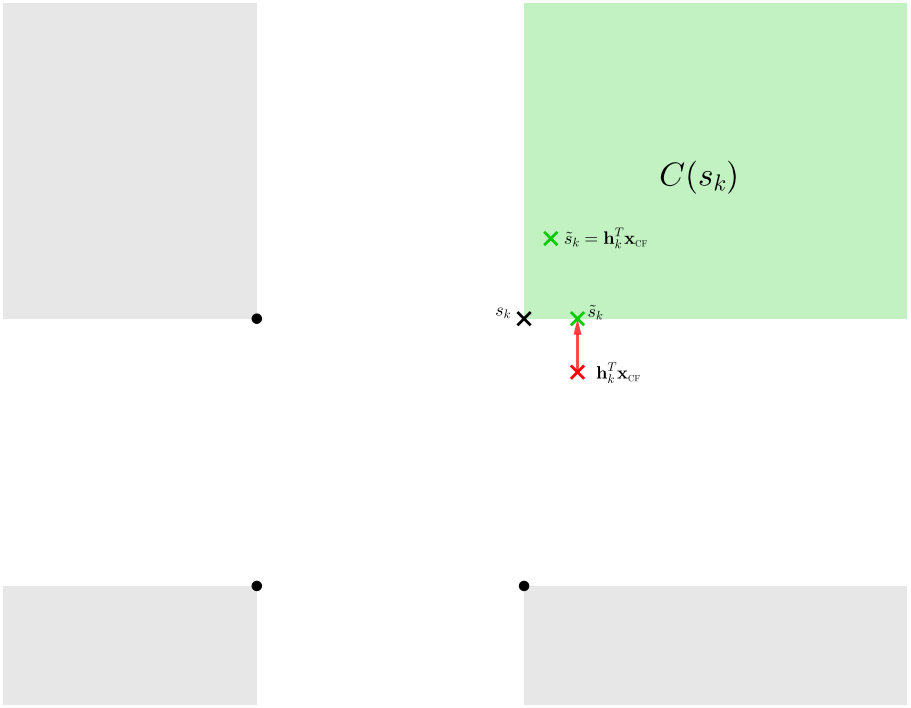

Figure 3.21: Illustration of projection of $\mathbf{h}_k^T \mathbf{x}_{\text{CF}}$ into $C(\mathbf{s})$ for QPSK.

meaning that ACE-BLS results in less peak regrowth than BLS. Since only the outermost constellation points are extended under ACE it is most effective for smaller constellations – where there is a higher probability of the received clipped signal falling within the allowed regions. This suggests that ACE is most applicable in scenarios with low to moderate individual user data rates.

When the number of antennas is reduced, the nullspace dimension, $M - K$, reduces, and the performance of the BLS algorithm reduces. The marginal improvement offered by using ACE is then increased, as shown in 3.23 for a reduced size MIMO system with $K = 8$ users and $M = 16$ antennas. Under QPSK signalling, ACE can provide over 2 dB of additional PAPR reduction, whilst achieving the same bit error rate, as shown in 3.24.

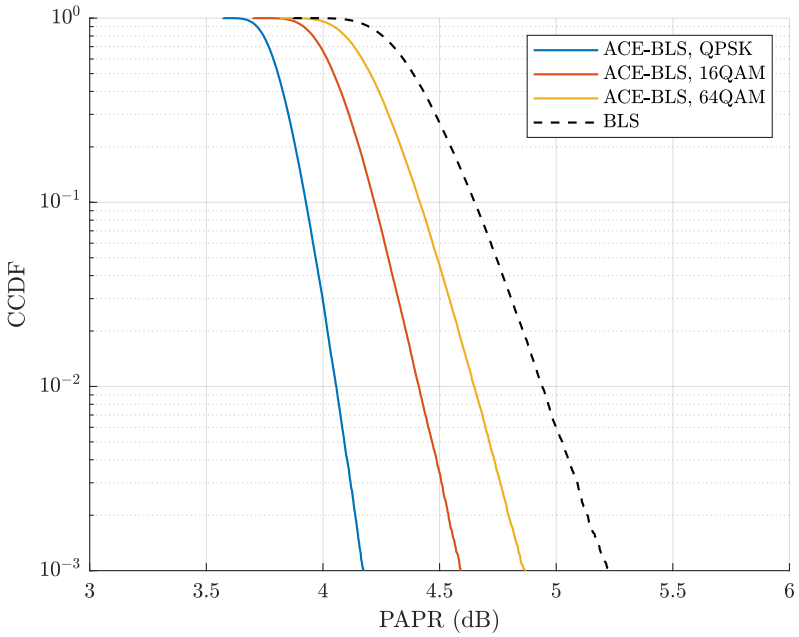

Figure 3.22: PAPR of ACE PAPR reduction scheme in 10 tap i.i.d Rayleigh channel. 5 clipping iterations, $\gamma = 1$, $K = 8$, $M = 64$, $N = 512$.

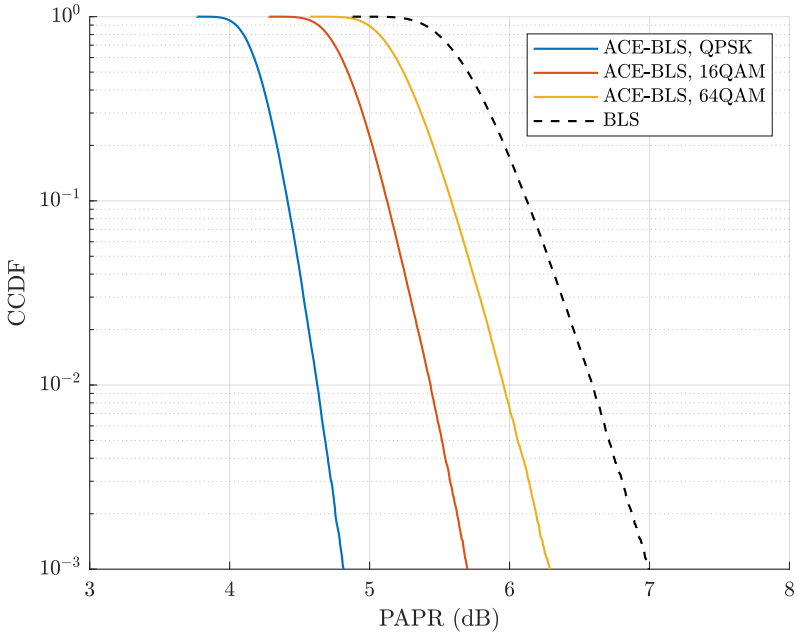

Figure 3.23: PAPR of ACE PAPR reduction scheme in 10 tap i.i.d Rayleigh channel. 5 clipping iterations, $\gamma = 1$, $K = 8$, $M = 16$, $N = 512$.

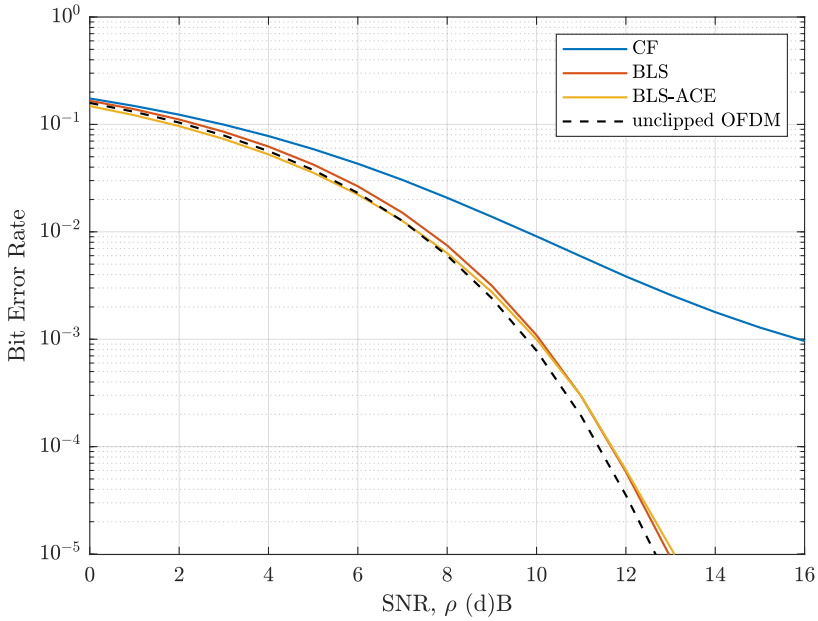

Figure 3.24: BER of QPSK under CF, BLS and ACE PAPR reduction schemes in 10 tap i.i.d Rayleigh channel. 5 clipping iterations, $\gamma = 1$, $K = 8$, $M = 16$, $N = 512$.

3.5.6 Practical Aspects

Finally, this section discusses some practical aspects of the proposed scheme: the computational complexity of the scheme is first explored and a basic algorithm for implementing the scheme is given. Finally, the implications of imperfect channel state information are discussed.

Computational Complexity

The key computations involved in the BLS scheme are:

- The FFTs and IFFTs required to convert the signal into the time domain for clipping, and back to the frequency domain. These each have complexity $\mathcal{O}\big(qN\log_2(qN)\big)$ per antenna, for a total complexity of $\mathcal{O}\big(qMN\log_2(qN)\big)$ per iteration.
- Calculation of the Bussgang matrix, $\mathbf{A}^{(i)}$. Using the expression in (3.97) this has overall complexity $\mathcal{O}\big(qMN\big)$ per iteration. From $\mathbf{A}^{(i)}$, the target symbol scalings μ_l can be calculated, with complexity $\mathcal{O}\big(qMN\big)$. These should be kept fixed for the duration of the coherence block (multiple subcarriers across multiple OFDM symbols), to ensure a constant symbol scaling is observed at the users.
- Calculation of the cancellation signal, requiring matrix multiplications for calculating the received clipped signal and the cancellation precoding, with overall complexity $\mathcal{O}\big(MKN\big)$. Since ZF precoding is optimal in many massive MIMO settings, it can be assumed that the ZF precoding matrix is available at no additional computational cost. Furthermore, existing signal processing software architecture used for data precoding could be re-used for calculating the LS signal.

For typical massive MIMO configurations ($K \ll M \ll N$), the overall complexity is

$$\text{computational complexity of BLS scheme} \sim \mathcal{O}\big(MN(q\log_2(qN)+K)\text{it}\big), \tag{3.104}$$

per OFDM symbol. This is linear in both MIMO dimensions and bandwidth, and hence scales well to large systems. Furthermore, overall complexity is comparable to that of conventional clipping & filtering – for example, in a system with 1200 subcarriers, oversampling factor 4, 64 BS antennas and 8 users the complexity of the BLS scheme is approximately double that of conventional CF per iteration. Since the PAPR reduction increases when more iterations are used, as shown in Figures 3.18 & 3.20, there is a tradeoff between performance and computational complexity.

Algorithm

The basic (unoptimised) BLS algorithm is shown in Algorithm 1.

Implications of Imperfect CSI

Practical MIMO systems will generally operate with imperfect CSI. Assuming the signal through the unknown channel component is treated as noise, two sources of interference will be produced

Algorithm 1 Bussgang-aware Least Squares Spatial Filtering

inputs: $\mathbf{x}[n], \mathbf{W}_l, \mathbf{H}_l, \mathbf{R}_x, \gamma$

$\mathbf{x}_{\text{BLS}}[n] \leftarrow \mathbf{x}[n]$

$A_{\max} \leftarrow \dfrac{\gamma}{M} \operatorname{Tr}\left(\mathbf{R}_x\right)$

$\mu_l \leftarrow 1$

$\mathbf{A} \leftarrow \mathbf{I}_M$

for $i = 1 : \text{it}$ **do**

  for $m = 1 : M$ **do**

    for $n = 0 : qN - 1$ **do**

      $x_{\text{C},m}[n] \leftarrow \min\left(A_{\max}, \left|x_{\text{BLS},m}[n]\right|\right) \times \frac{x_{\text{BLS},m}[n]}{|x_{\text{BLS},m}[n]|}$

    end for

    $[\mathbf{A}]_{m,m} \leftarrow \frac{\sum_{n=0}^{qN-1} x_{\text{C},m}^{(i)}[n] x_{\text{BLS},m}^{(i-1)*}[n]}{\sum_{n=0}^{qN-1} |x_{\text{BLS},m}^{(i-1)}[n]|^2}$

  end for

  $\mathbf{x}_{\text{C},l} \leftarrow \text{FFT}\big(\mathbf{x}_{\text{C}}[n]\big)$

  for $l = 0 : N - 1$ **do**

    $\mu_l = \frac{\text{Tr}(\mathbf{W}_l^\dagger \mathbf{A} \mathbf{W}_l)}{\text{Tr}(\mathbf{W}_l^\dagger \mathbf{W}_l)} \mu_l$

    $\mathbf{x}_{\text{BLS},l} \leftarrow \mathbf{x}_{\text{C},l} - \mathbf{W}_l\big(\mathbf{H}_l \mathbf{x}_{\text{C},l} - \mu_l \mathbf{x}_l\big)$

  end for

  $\mathbf{x}_{\text{BLS}}[n] \leftarrow \text{IFFT}\big(\mathbf{x}_{\text{BLS},l}\big)$

end for

$\mathbf{x}_{\text{BLS}}[n] \leftarrow \zeta \mathbf{x}_{\text{BLS}}[n]$

outputs: $\mathbf{x}_{\text{BLS}}[n]$

under the BLS scheme – interference that results from the data symbol precoding not perfectly matching the channel, and interference that results from the clipping noise not being perfectly cancelled out by the BLS method.

Using the analysis method outlined in Section 2.4.2, the received signal under the BLS scheme is given by

$$y_k = \hat{\mathbf{h}}_k^T \mathbf{x}_{\text{BLS}} + \mathbf{e}_k^T \mathbf{x}_{\text{BLS}} + \eta \tag{3.105}$$

$$= \hat{\mathbf{h}}_k^T \mathbf{x}_{\text{BLS}} + \mu \mathbf{e}_k^T \mathbf{W} \mathbf{s} + \mathbf{e}_k^T \big(\mathbf{I}_M - \mathbf{W}^{(\text{ZF})} \mathbf{H}\big) \boldsymbol{\varepsilon} + \eta \tag{3.106}$$

$$= \mu \hat{\mathbf{h}}_k^T \mathbf{w}_k s_k + (\mu \sum_{j \neq k} \hat{\mathbf{h}}_k^T \mathbf{w}_j s_j + \theta_k + \phi_k) + \eta \tag{3.107}$$

$$= \text{signal} + \text{interference} + \text{noise} \tag{3.108}$$

where all terms are uncorrelated and θ_k is interference due to data precoding and ϕ_k interference due to uncancelled clipping noise. Since the power of the data signal component is much larger

than the power of the clipping noise it can generally be assumed that $\mathbb{E}\big[\big|\theta_k\big|^2\big] \gg \mathbb{E}\big[|\phi_k|^2\big]$, and the data precoding is the main limiting factor, rather than the BLS scheme.

Some insights can be gained from assuming the clipping analysis model in Section 2.4.2, under which

$$\mathbb{E}\big[\big|\theta_k\big|^2\big] = \alpha^2\rho\sum_{j=1}^{K}\mathbf{w}_j^\dagger\mathbf{C}_k\mathbf{w}_j \tag{3.109}$$

$$\mathbb{E}\big[|\phi_k|^2\big] = \sigma_\varepsilon^2\,\mathrm{Tr}\Big(\big(\mathbf{I}_M - \mathbf{W}^{(\mathrm{ZF})}\mathbf{H}\big)\mathbf{C}_k\Big) \tag{3.110}$$

$$< \sigma_\varepsilon^2\,\mathrm{Tr}\big(\mathbf{C}_k\big), \tag{3.111}$$

where the inequality follows since $\big(\mathbf{I}_M - \mathbf{W}^{(\mathrm{ZF})}\mathbf{H}\big)$ is a nullspace projection. In comparison, under standard clipping & filtering

$$y_k = \hat{\mathbf{h}}_k^T\mathbf{x}_{\mathrm{CF}} + \mathbf{e}_k^T\mathbf{x}_{\mathrm{CF}} + \eta \tag{3.112}$$

$$= \hat{\mathbf{h}}_k^T\mathbf{A}\mathbf{w}_k s_k + \big(\sum_{j\neq k}\hat{\mathbf{h}}_k^T\mathbf{A}\mathbf{w}_j s_j + \hat{\mathbf{h}}_k^T\boldsymbol{\varepsilon} + \mathbf{e}_k^T\mathbf{A}\mathbf{W}\mathbf{s} + \mathbf{e}_k^T\boldsymbol{\varepsilon}\big) + \eta \tag{3.113}$$

$$= \alpha\hat{\mathbf{h}}_k^T\mathbf{w}_k s_k + \big(\alpha\sum_{j\neq k}\hat{\mathbf{h}}_k^T\mathbf{w}_j s_j + \hat{\mathbf{h}}_k^T\boldsymbol{\varepsilon} + \theta_k + \phi_k'\big) + \eta \tag{3.114}$$

$$= \text{signal} + \text{interference} + \text{noise}. \tag{3.115}$$

Under the simplified clipping analysis model

$$\mathbb{E}\big[|\hat{\mathbf{h}}_k^T\boldsymbol{\varepsilon}|^2\big] = \sigma_\varepsilon^2\big\|\hat{\mathbf{h}}_k\big\|^2 \tag{3.116}$$

$$\mathbb{E}\big[|\phi_k'|^2\big] = \sigma_\varepsilon^2\,\mathrm{Tr}\big(\mathbf{C}_k\big). \tag{3.117}$$

From this it can be seen that BLS scheme experiences lower amounts of interference from clipping noise compared to standard clipping & filtering, and better SINR, for any quality of CSI. However, as the CSI error grows large the error due to data precoding with a poor estimate of the channel matrix, θ_k, will come to dominate over all of the clipping noise effects. At this point, the benefits of using BLS will be small compared to just using standard clipping & filtering.

Generally, it can be expected that the BLS scheme will experience improved SINR over conventional clipping & filtering for any quality of CSI, since the error cancellation stage will always cancel out some of the error introduced by clipping & filtering. However, when the CSI is of poor quality, performance degradation will be mainly due to imperfect CSI, rather than clipping noise, and the BLS scheme will not significantly improve performance. A thorough numerical investigation into the benefits of the scheme under imperfect CSI remains as future work.

3.6 Conclusion

The high PAPR of OFDM signals is a long-standing problem that affects the energy efficiency and hardware cost of cellular systems. The precoded downlink signals used in massive MIMO-

OFDM systems suffer from worse PAPR than SISO systems, and therefore new practical and effective PAPR reduction schemes are required to help massive MIMO achieve the high energy efficiencies required by fifth generation & future wireless systems.

The primary contribution of this chapter is a clipping & spatial filtering PAPR reduction scheme that uses clipping to reduce the PAPR of the transmit signal, simultaneously transmitting an additional low power spatially precoded signal to correct the distortion introduced at the receivers by the clipping operation. This enables aggressive levels of clipping to be applied to the BS transmit signal without incurring the significant performance degradation that occurs under conventional clipping & filtering. Numerical results show that for a massive MIMO-OFDM system with 8 users, 64 BS antennas and 512 subcarriers, the scheme can reduce the PAPR of the transmit signal by over 8 dB whilst only incurring a performance loss of 0.3 dB.

Whilst related clipping & spatial filtering schemes have previously been proposed, the method developed here explicitly accounts for the signal attenuation caused by the clipping operation, improving convergence at low clipping ratios to facilitate 1 dB additional PAPR reduction compared to previous schemes. This improvement comes from developing & incorporating a statistical model for the clipping operation, based on Bussgang's theory, that decomposes the clipped signal into a linear transformation (attenuation) of the transmit signal and an uncorrelated clipping noise.

The proposed solution is extended to include active constellation extension, where certain distortion is allowed at the users provided it falls outside of the signal constellation. This enables a further 1-2 dB PAPR reduction with no additional performance cost, and is shown to be particularly beneficial when smaller constellation sizes are used and/or when a smaller excess of antennas is used at the BS. It could therefore be a valuable addition to massive MIMO systems that simultaneously serve low to medium data rates to a large numbers of users.

Overall, the proposed solution represents an approach to PAPR reduction for massive MIMO which is both effective and practical. The scheme can be used 'on-top' of a conventional linearly precoded massive MIMO-OFDM system, and its main computational tasks re-use the conventional FFT and linear ZF precoding operations already used within MIMO-OFDM systems. For a typical system the computational complexity is around double that of conventional clipping & filtering, representing an attractive compromise between performance and computational requirement.

However, there may be some scenarios where this additional complexity is not justified – for example, when operating at low SNR or with poor quality CSI the relative impact of clipping noise on performance is reduced, and the use of conventional clipping & filtering without spatial filtering may give adequate performance. Further work to characterise the benefits under different operating conditions would therefore be a useful next step in the development of the scheme.

The findings of this chapter suggest that another standout area requiring investigation is the impact of precoding and channel conditions on the signal PAPR. Results presented here show, for example, that precoding for a correlated fading channel can result in a MIMO-OFDM signal with 2+ dB higher PAPR than under i.i.d fading. Thus, whilst the proposed solution achieves similar levels of PAPR *reduction* in both the i.i.d and correlated fading examples, the

overall PAPR remains higher under correlated fading. One factor that should be investigated is the influence on PAPR of an unequal power distribution between the transmit antennas, since a spread of powers brings an inherent PAPR increase. Conventional MIMO precoding matrices are designed under total power constraints, but results presented in this chapter indicate that designing the precoding matrices using per-antenna power constraints can significantly reduce PAPR. Low complexity precoding schemes that minimise power variations between antennas could therefore complement the proposed solution to bring further PAPR reduction.

Chapter 4

Transform Coding-based Signal Compression for Uplink MIMO C-RAN

The MIMO cloud (or centralised) radio access network (C-RAN) represents a promising concept for handling the densification of cellular networks required to meet future capacity demands [168]. Performing the processing for multiple geographically distributed remote radio heads at a single central processor (CP) enables inter-user interference to be eliminated and good service provided to large numbers of users within the macro-coverage area. The computationally intensive signal processing, MAC and higher layer network functions are performed in software at the CP, and upgrades to the network require only CP software updates and the deployment of additional low-cost radio heads. Furthermore, distributing the radio heads reduces the path-loss to users, enabling transmit power to be reduced and energy efficiency improved.

In the C-RAN architecture a 'fronthaul' network provides the connections between the radio heads and CP, and ideally consists of high capacity fibre links. However, in many scenarios the provision of a full fibre connection to each radio head is prohibitively expensive, leading to a growing interest in the use of finite-capacity fronthaul based on reduced capacity fibre, ethernet or wireless point-to-point links[1] [169]. Under the rudimentary signal compression techniques that exist in current cellular systems, these finite-capacity fronthaul connections can become a bottleneck for system performance, severely constraining the sum capacity of the C-RAN network [98]. This motivates the development of bespoke signal compression schemes targeted towards optimising the performance of a C-RAN network with finite-capacity fronthaul links.

Focusing on the compression of uplink signals, this chapter investigates the use of transform coding for compression of the received signals in MIMO systems with multi-antenna receivers, with particular focus on the operating region in which the system capacity is limited by the capacity of the fronthaul network. Under the proposed transform coding scheme, a lossless linear transformation is applied to the correlated received signal vector at each remote receiver, before the components of the transformed signals are individually compressed, at appropriate resolutions, using scalar compression, as illustrated in Figure 4.1. Assuming optimal Gaussian

[1] In 2016, around 40% of connections used wireless links [76]

scalar compression (an information theoretic model that provides an tractable approximation of entropy coded scalar compression), expressions are derived for sum and per-user network capacities under transform coding, and performance optimised through the choice of linear transformations and scalar compression rates. A 'fronthaul efficiency' metric is introduced as the ratio of total uplink MIMO throughput/capacity to total fronthaul throughput/capacity, to quantify how well the compression scheme utilises the limited fronthaul capacity.

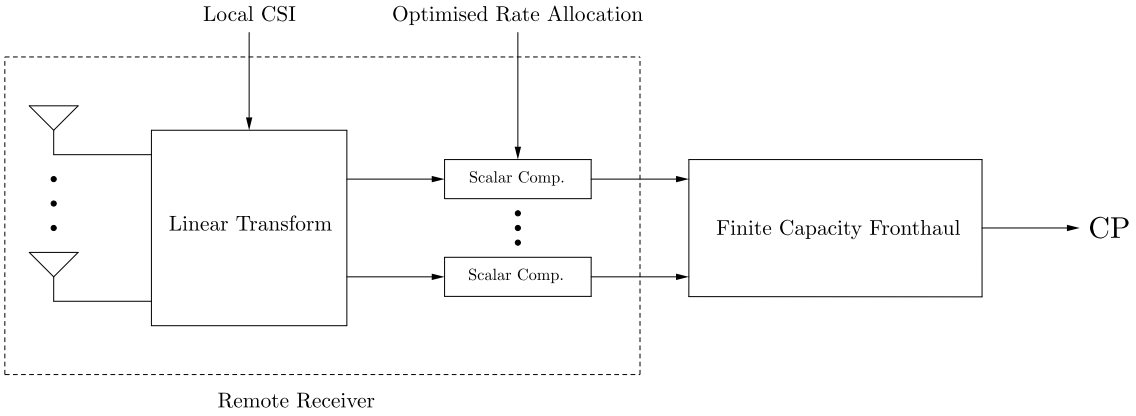

Figure 4.1: Block diagram of proposed transform coding fronthaul compression scheme.

First, the compression of massive MIMO uplink signals in a system with a single remote receiver is considered. This is relevant to a scenario where the processing for different massive MIMO receivers is performed separately but at the same CP, and serves as a good starting point for analysing the use of transform coding in MIMO systems. A new sum capacity upper bound for the case where the signals at each antenna are individually compressed/quantized is first developed, showing that this leads to very poor fronthaul efficiency when the number of antennas grows large. It is then shown that this issue is overcome by exploiting the underlying *sparsity* in the massive MIMO received signal through transform coding.

The sum capacity maximising transform coding scheme for a single MIMO receiver is known to use the Karhunen-Loeve transform (KLT) in a conjunction with a waterfilling rate allocation [226]. Here, this is applied to massive MIMO systems, and shown to use the available fronthaul very efficiently, particularly at high SNR. A separate transform coding scheme that produces a uniform quantization noise level (UQN) across the signal is then shown to approximate the waterfilling allocation at higher fronthaul rates, and shown numerically to achieve similar performance at all rates. Under this UQN scheme it is shown, using a combination of mathematical arguments and numerical results, that, whilst the MIMO sum capacity is limited by the amount of fronthaul available, some of the asymptotic benefits of deploying large numbers of antennas are preserved – linear detection is optimal, fast fading disappears and required transmit power reduces.

The remainder of the chapter focuses on transform coding compression for distributed MIMO, where the CP uses the individually transform coded signals from multiple receivers to *jointly* detect the user symbols. Whilst this is known to be the optimal point-to-point (P2P) compression strategy for distributed MIMO, in contrast to the single receiver case no closed

form solutions for finding the sum capacity maximising transforms and rate allocations exist. Instead, these must be found using a successive convex approximation approach (SCA-P2P) in which a sequence of convex optimisations is solved using computationally expensive numerical solvers [248] – a procedure that is not practical to implement in real distributed MIMO C-RAN deployments.

First, the performance of local transform coding with a UQN rate allocation is analysed in the distributed MIMO setting. Whilst this has been shown to be approximately optimal at high fronthaul rates, a simple high SNR sum capacity upper bound is derived that shows that when the network has an overall excess of antennas and limited fronthaul UQN compression will tend to utilise the available fronthaul poorly. The benefits of using the SCA-P2P approach are then demonstrated using numerical examples.

To address the respective issues of performance and computational complexity of the UQN and SCA-P2P schemes, a third approach is investigated here in which a *fixed transform* is used in conjunction with a *jointly optimised rate allocation*. Using the local KLT as the transform at each receiver, this scheme is an adaptation of the single receiver transform coding scheme in which the local rate allocations are replaced by ones that are jointly calculated for all receivers using global CSI.

A similar concept was previously explored in [123], based on maximising the minimum user capacity under fixed rate scalar quantization. Here, a new rate allocation scheme is proposed that uses a successive convex approximation approach to maximise the sum capacity, under either linear or non-linear symbol detection, with optimal Gaussian scalar compression (SCA-RA). This is shown to have a simple iterative solution in which a waterfilling-type rate allocation is performed and the corresponding MMSE detection matrices calculated. Unlike the approaches in [123] this does not require the use of any numerical solvers, with numerical results indicating convergence to an acceptable solution requires only a small number of iterations (~ 5).

Overall, the proposed SCA-RA scheme is shown using numerical examples to incur only a small performance loss compared to the optimal SCA-P2P transform coding solution, and greatly outperform local transform coding with UQN rate allocation. Analysis of the rate allocation scheme shows that when the system has an overall excess of antennas, the proposed scheme will tend to produce to a sparse rate allocation when operating in the fronthaul-limited region – in response to the joint sparsity intrinsically present in the network. Numerical results show that the scheme performs best at high SNR, continues to achieve good fronthaul efficiently as the network densifies, and that there is a benefit to deploying additional antennas at each receiver. The reduction in signalling overheads compared to the SCA-P2P scheme are also discussed.

The transform coding schemes for both massive MIMO and distributed MIMO are adapted for the case of imperfect CSI, and the importance of having good CSI estimates is shown.

Overall, transform coding with the KLT transform and optimised rate allocations represents a scalable solution for exploiting sparsity and dependencies in the received signals to achieve efficient fronthaul compression in both massive MIMO and distributed MIMO C-RAN.

4.1 Chapter Overview

The chapter has the following general structure:

- Section 4.2 provides the background to the C-RAN fronthaul compression problem. It begins with a more through exposition of the need for bespoke signal compression scheme for C-RAN systems. A brief summary of lossy signal compression theory is provided, before a review of the state-of-the-art fronthaul compression techniques for MIMO C-RAN is provided.

- Section 4.3 studies the application of transform coding to massive MIMO. It begins by considering the direct scalar compression of the received signals, before outlining the optimal and UQN transform coding schemes. The performance of massive MIMO systems under UQN transform coding is then analysed, and the impact of increasing the number of BS antennas discussed. The method is then adapted for the case of imperfect CSI at the receiver.

- Section 4.4 studies the application of transform coding to distributed MIMO. A transform scheme that jointly optimises the compression rates at each receiver is derived, and shown to deliver a good compromise between complexity and performance. Numerical results are provided to analyse its performance under different conditions, before the scheme is adapted for imperfect CSI and practical aspects discussed.

- Section 4.5 summarises the findings and provides some concluding remarks and directions for future work.

4.1.1 Novel Contributions

The key contributions to the state-of-the-art made in this chapter are:

- **A simple upper bound on the sum capacity achievable under direct sampling & forward, Section 4.3.1**. For the case where the received signals at each antenna are individually compressed, it is shown that the achievable sum capacity increases slowly with the available fronthaul capacity when the BS has a large excess of receive antennas, resulting in poor fronthaul efficiency.

- **Showing that for a single MIMO receiver, transform coding using the KLT and Gaussian scalar compression asymptotically achieves the cut-set bound at high SNR, Section 4.3.2.** This demonstrates that the compress & forward strategy with transform coding can achieve high fronthaul efficiency, at the expense of increased transmit power and reduced energy efficiency.

- **Demonstrating the benefits of adding more antennas under favourable propagation conditions, Section 4.3.4.** This follows first from showing that with a single receiver, transform coding with the KLT and a uniform quantization noise level is quasi-optimal, from which it is then shown that a number of the asymptotic massive MIMO

results still hold when the performance is limited by the fronthaul capacity – linear processing becomes optimal, fast fading disappears and user transmit power reduces.

- **A scalable transform coding scheme for distributed MIMO that uses jointly optimised rate allocations to efficiently compress the received signals at different receivers, Section 4.4.3.** Whilst the approach taken is similar to previously used in and [123], here a lower complexity scheme is developed that doesn't require numerical solvers. This comes close to the performance of optimal point-to-point compression, and at high SNR can achieve very good fronthaul efficiency in the fronthaul-limited region.
- **Adapting both schemes for the case of imperfect CSI, sections 4.3.5, 4.4.5.** Assuming MMSE channel estimation is used, both schemes are adapted for the case of imperfect CSI at the receivers, and capacity expressions derived. The importance of having good CSI estimates is demonstrated.

4.1.2 Published Work

The basis of the transform coding approach was published in [223]. This applied transform coding with the KLT to the MIMO uplink under Gaussian scalar compression, and derived capacity expressions for massive MIMO and distributed MIMO scenarios. The idea of using local and joint rate allocations were discussed, and numerical results provided.

However, various key results provided in this chapter remain unpublished. Notably, the analysis of massive MIMO under UQN compression and results on increasing the number of antennas under favourable propagation are new, whilst the joint rate allocation in [223] used a simple non-linear gradient descent, as opposed to the more rigorously derived successive convex approximation approach here. All of the results for transform coding with imperfect CSI are unpublished. It is anticipated that some combination of these findings could form the basis of at least one additional conference paper.

4.2 Background

This section outlines the motivation behind the research in this chapter and Chapter 5, as well as summarising the relevant existing research efforts.

4.2.1 Motivation

The simplest radio head receiver would consist of just an RF frontend and ADC – downconverting and digitising the received signals and transferring the raw IQ samples back to the CP for processing. However, in scenarios where the capacity of the fronthaul connections are constrained, such a system is not always feasible due to the large amounts of data that are produced. Consider a MIMO C-RAN uplink system where a remote receiver equipped with M antennas is connected via a fronthaul connection to a CP. If the system has operating bandwidth B Hz, and the baseband signals at each antenna are sampled at a resolution of n_b bits (per I & Q complex dimension), with an oversampling factor q, the total bit rate of the sampled signals

at the receiver is

$$\text{total sampled bit rate} = M \times B \times q \times 2n_b \quad \text{bps.} \tag{4.1}$$

A system with bandwidth 100 MHz, oversampling factor 1.2, and 15 bit resolution [44] generates 3.6 Gbps of data per antenna, and as the number of antennas increases the quantity of sampled data becomes very large – 16 antennas producing over 50 Gbps.

The amount of actual user uplink data conveyed by the received signals,

$$\text{total uplink user throughput} = \text{SSE} \times B \quad \text{bps,} \tag{4.2}$$

is considerably lower than this. For example, a MIMO system with average sum spectral efficiency of 40 bps/Hz would achieve a total throughput of 4 Gbps in a 100 MHz channel.

Transferring the raw IQ samples across the fronthaul for MIMO processing and symbol decoding at the CP is therefore very *fronthaul inefficient* – the fronthaul capacity required is orders of magnitude larger than the quantity of useful data conveyed,

$$\text{fronthaul efficiency} = \frac{\text{total uplink user throughput}}{\text{total fronthaul throughput}}. \tag{4.3}$$

In deployments where fronthaul connections are provided by dedicated high capacity optical fibre a poor fronthaul efficiency may not be an issue. However, as cellular systems densify, the capital expenditure associated with deploying dedicated fibre fronthaul for each radio head becomes prohibitive. This has led to cheaper fronthaul, more flexible, solutions based on either shared/lower grade fibre [168], or wireless point-to-point links [169] being considered. Recently, there has been a particular growing interest in the use of mmWave bands (above 60 GHz) for providing fronthaul, exploiting the underutilised spectrum at higher frequencies to provide multi-gigabit fronthaul capacities of between 1 and 10 Gbps [98].

A potential solution for limited fronthaul systems is to perform the full MIMO detection and symbol decoding at the remote receiver, and transfer the decoded user data streams over the fronthaul – a strategy known as 'decode & forward'. However, this approach comes at the expense of increased radio head complexity and therefore higher cost. Furthermore, in a distributed MIMO C-RAN system, optimal MIMO detection requires joint processing of all received signals, and local decoding therefore generally cannot capture the full benefits of the C-RAN architecture.

The splitting of functionality between the remote receivers and CP has been an area of practical research [52], [129]. Performing some simple parts of the receive signal processing chain at the remote receiver, such as the IFFT part of OFDM processing, can enable redundant temporal features to be removed from the received signal, and significantly reduce the amount of data that must be transferred over fronthaul [44]. However, for the best utilisation of the fronthaul it is necessary to also account for spatial dependencies between the received signals, both locally (between different antennas) and across the C-RAN network (between different receivers).

From an information theory perspective, the study of fronthaul constrained systems has roots in network information theory, a relatively recent branch of information theory dedicated

to studying the communication limits of networks of interconnected nodes. The fronthaul constrained C-RAN networks considered in this thesis are two-hop networks, in which the first hops – between the users and remote receivers – are wireless channels, whilst the second hops – the fronthaul connections between remote receivers and CP – are treated as noiseless links that can achieve error-free communication up to a certain rate. This fronthaul model, shown in Figure 4.2, is appropriate for either rate limited wired/fibre connections or point-to-point wireless links (which have a fixed, often line-of-sight, channel and thus do not experience significant time dependent fading).

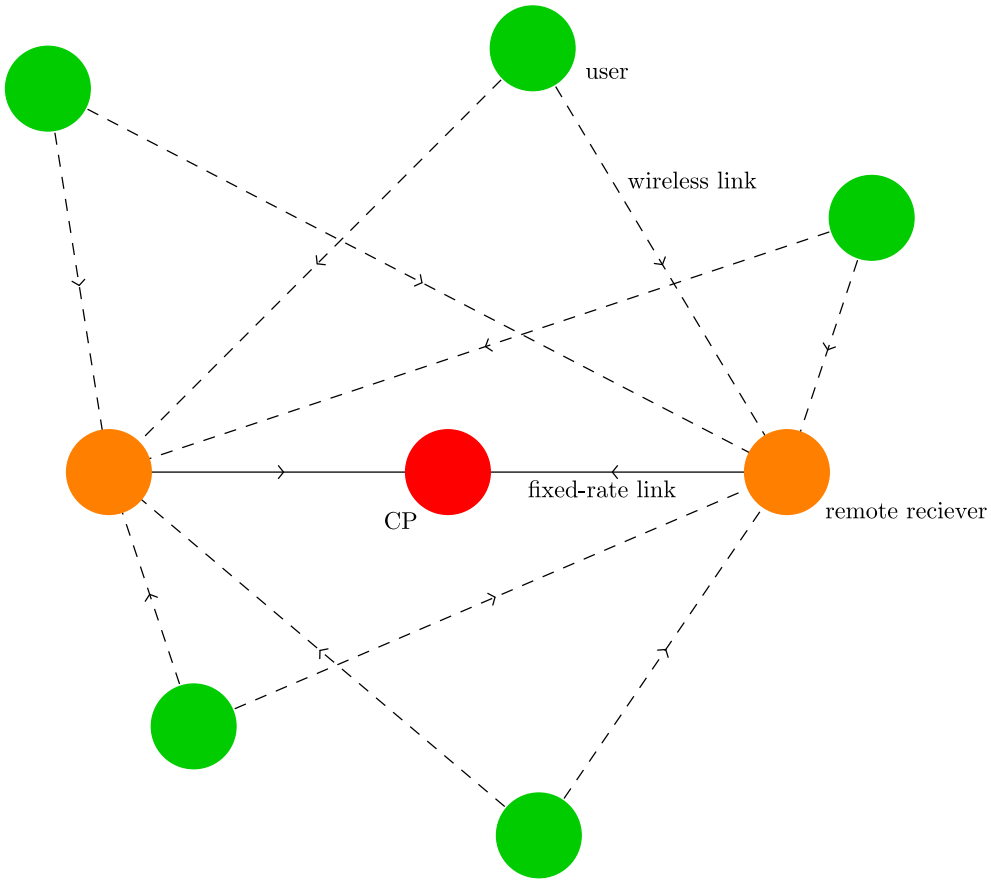

Figure 4.2: Fronthaul constrained C-RAN network topology with $K = 5$ users, $L = 2$ receivers and central processor.

Potential signalling strategies for this network include:

- Sample & forward, in which the receivers simply digitize their received signals and forward the sampled signals over fronthaul to the CP.
- Decode & forward, or partial decode & forward, in which the receivers either fully or partially decode the users messages, and forward them over fronthaul to the CP.
- Compress & forward, in which the receivers compress their received signals, and then forward the compressed signals over fronthaul to the CP, where they are decoded.

The research in this thesis focuses on the use of compress & forward, in which the CP receives compressed signals over the fronthaul connections and performs MIMO detection jointly on

the decompressed signals. This approach treats the fronthaul as a black box and the artefacts introduced by compression as a noise, and can be used with conventional transmission techniques and standard MIMO detection methods. It is flexible, and can easily be modified to incorporate different topologies – e.g. the scenario where some fronthaul links have finite capacity and others infinite. Throughout this work, it is assumed that OFDM signalling is used, and that all compression is performed on a subcarrier level, with IFFT processing first carried out at the receiver. There is a particular focus on low complexity methods that exploit the spatial characteristics of the received signals in large-scale MIMO systems.

4.2.2 Fundamentals of Signal Compression

Lossy signal compression addresses the problem of encoding the 'best' representation, $\tilde{\mathbf{y}}$, of a continuously valued signal, $\mathbf{y}$, using a finite number of bits. As part of his groundbreaking information theory work [191], Shannon established the theoretical basis of lossy signal compression, showing that the minimum average number of bits, $\mathcal{R}$, required to encode a lossy compressed signal is equal to the mutual information between the original and compressed signals

$$\mathcal{R} = \mathcal{I}\big(\tilde{\mathbf{y}}; \mathbf{y}\big). \tag{4.4}$$

His work addressed the problem of establishing a relationship between the encoding rate of a compression scheme, and the distortion it introduced. In his rate-distortion theory, distortion, $\mathcal{D}$, is quantified by some metric that measures the difference between the original signal and the compressed signal, such as mean squared error,

$$\mathcal{D} = \mathbb{E}\big[\|\tilde{\mathbf{y}} - \mathbf{y}\|^2\big], \tag{4.5}$$

and a rate-distortion function, $\mathcal{D}(\mathcal{R})$, is sought. The ideas from rate-distortion theory have been hugely influential, and form the basis of modern image, audio and video compression schemes.

In the fronthaul compress and forward scheme, compression is applied to the received signals at each remote radio head, before they are forwarded over fronthaul to the CP for user symbol detection. The fronthaul compression and rate-distortion problems are therefore distinct: the latter aims to produce an accurate representation of the received signal, whilst the former aims to produce a signal representation from which the underlying user symbols can be reliably recovered. As a result, it is possible for a fronthaul compression scheme to produce compressed signals that are a poor representation of the original received signals (in terms of distortion), but which achieve good MIMO symbol detection performance. Conversely, compressing the signals to minimise distortion can lead to poor overall MIMO performance. There is thus a need for bespoke compress-and-forward solutions for fronthaul constrained C-RAN.

This section briefly outlines the fundamentals of scalar and vector point-to-point compression, before briefly discussing the idea of distributed compression. As with the capacity analysis elsewhere in this thesis, the treatment assumes that all symbols are drawn from a Gaussian distribution.

Gaussian Scalar Compression

Optimal compression of a complex Gaussian scalar, $y \sim \mathcal{CN}(0, \sigma^2)$, is characterised by the 'Gaussian test channel[2]',

$$\tilde{y} = y + \delta, \tag{4.6}$$

where $\delta \sim \mathcal{CN}(0, \phi)$ is Gaussian quantization noise that is independent of y. The minimum average encoding rate (bits), r, is

$$r = \mathcal{I}(\tilde{y}; y) \tag{4.7}$$

$$= \log_2\left(1 + \frac{\sigma^2}{\phi}\right) \tag{4.8}$$

Rearranging, the quantization noise variance therefore scales as

$$\phi = \frac{\sigma^2}{2^r - 1} \tag{4.9}$$

$$\approx \sigma^2 2^{-r}, \tag{4.10}$$

where the approximation is tight for $r > 2$, as shown in Figure 4.3. When $r = 0$, $\phi = \infty$, accurately modelling the scenario where y is not encoded.

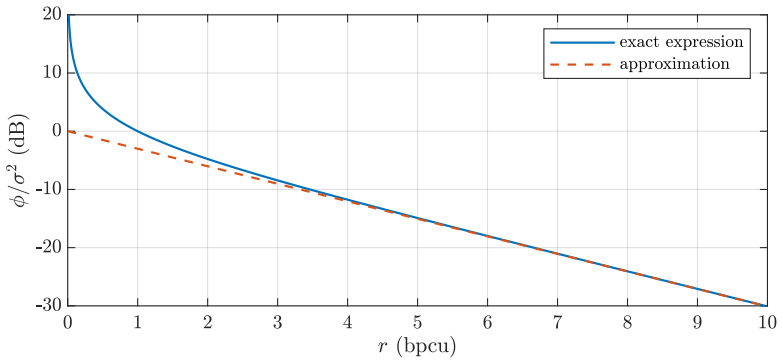

Figure 4.3: Exact and approximate quantization noise variance.

The Gaussian test channel model is an information-theoretic concept, and cannot be perfectly realised in practice. However, block coding compression schemes based on trellis coding [131] and sphere coding [77] have been shown to achieve comparable performance with manageable complexity, whilst at high rates simple uniform quantization followed by entropy coding incurs a rate penalty of only 0.5 bits [69]. The simplicity of the test channel model therefore makes it useful for analysing performance of systems under compression, and has been widely used.

In [227] it is shown that when the signal being compressed takes the form of a scalar AWGN channel,

$$y = hx + \eta, \tag{4.11}$$

[2]This is known as the forward test channel. Gaussian compression can also be modelled by a reverse test channel, where δ is instead independent of $\tilde{y}$. For analysis purposes, the forward test channel is most appropriate here.

the Gaussian test channel also maximises the mutual information between $\tilde{y}$ and x,

$$\mathcal{I}(\tilde{y};x) = \log_2\Big(1+\frac{\rho|h|^2}{\phi+1}\Big) \tag{4.12}$$

$$= \log_2\Big(1+(2^r-1)\frac{\rho|h|^2}{\rho|h|^2+2^r}\Big), \tag{4.13}$$

which is the capacity of the compressed and forwarded scalar AWGN channel, compressed at average rate r bits per channel use.

Fixed-Rate Scalar Quantization

In many practical applications, fixed-rate quantization – in which each quantized sample is encoded individually as a string of b bits – is preferred to block coding methods, because of its simplicity. For $y \sim \mathcal{CN}(0,\sigma^2)$, optimal fixed-rate quantization is achieved by individually quantizing the real and imaginary components,

$$\tilde{y} = Q_c(y) \tag{4.14}$$

$$= Q\big(\Re(y)\big) + jQ\big(\Im(y)\big), \tag{4.15}$$

using quantization points found using the Lloyd-Max algorithm (and stored in a look-up table).

The output of the fixed-rate quantizer can be expressed in a similar form to (4.6),

$$\tilde{y} = y + \delta. \tag{4.16}$$

where δ is quantization noise with $\mathbb{E}[|\delta|^2] = \phi$ but *arbitrary* distribution. For $b \geq 6$ it can be shown that

$$\phi \approx \sigma^2 \frac{\pi\sqrt{3}}{2} 2^{-b}, \tag{4.17}$$

and comparing to (4.9), an additional $\log_2\left(\pi\sqrt{3}/2\right) \approx 1.4$ bits are required compared to optimal Gaussian scalar compression [72], with the additional constraint that the number of bits be a multiple of two (integer for both real and imaginary), $b \in 2\mathbb{Z}^+$. Since Gaussian noise is the worst case noise distribution from a capacity perspective [84], analysis that uses the Gaussian test channel can be easily generalised to fixed-rate quantization.

Gaussian Vector Compression

Compression of a complex Gaussian vector, $\mathbf{y}$, can also be modelled using a test channel,

$$\tilde{\mathbf{y}} = \mathbf{y} + \boldsymbol{\psi}, \tag{4.18}$$

where $\boldsymbol{\psi} \sim \mathcal{CN}\big(0,\boldsymbol{\Psi}\big)$, with $\boldsymbol{\Psi}$ a suitably chosen covariance matrix. Using the eigendecomposition

$$\boldsymbol{\Psi} = \mathbf{F}\boldsymbol{\Phi}\mathbf{F}^\dagger, \tag{4.19}$$

the compressed signal can be transformed into

$$\tilde{\mathbf{z}} = \mathbf{F}^{\dagger}\tilde{\mathbf{y}} \tag{4.20}$$

$$= \mathbf{F}^{\dagger}\mathbf{y} + \boldsymbol{\delta}, \tag{4.21}$$

where $\boldsymbol{\delta} \sim \mathcal{CN}(0, \boldsymbol{\Phi})$ is spatially white quantization noise (since $\boldsymbol{\Phi}$ is diagonal). For a given $\boldsymbol{\Psi}$, the vector test channel can therefore be implemented using a *transform coding* approach:

- Apply the linear unitary transform $\mathbf{F}^{\dagger}$ to $\mathbf{y}$, to produce the transformed vector $\mathbf{z}$,

$$\mathbf{z} = \mathbf{F}^{\dagger}\mathbf{y}. \tag{4.22}$$

- Independently compress the elements of $\mathbf{z}$ using M scalar compression codebooks with rates calculated using (4.7) to give the desired quantization noise variances $\boldsymbol{\Phi}$.

- Reverse transform $\tilde{\mathbf{z}}$ to produce $\tilde{\mathbf{y}}$,

$$\tilde{\mathbf{y}} = \mathbf{F}\tilde{\mathbf{z}}. \tag{4.23}$$

When the compressed signal is used for MIMO detection the final reverse transform stage is in fact unnecessary, since MIMO processing can be equivalently performed directly on $\tilde{\mathbf{z}}$.

An interesting observation made in [226] is that the vector compression strategies for minimising distortion in $\tilde{\mathbf{y}}$ and for maximising the mutual information $\mathcal{I}\big(\tilde{\mathbf{y}}; \mathbf{x}\big)$ are different – they use the same linear transform, but different rate allocations. This result is discussed in more detail in the following section, which considers transform coding signal compression for massive MIMO systems.

When $\mathbf{y}$ is non-Gaussian, the transform coding approach is generally sub-optimal, and vector quantization methods [73] can give improved performance. However, the optimal vector quantization regions depend on the statistical distribution of the signal and must be calculated numerically. When the distribution is unknown, or varies with time (such as for a time varying MIMO channel), this becomes impractical, and hence transform coding is widely used in many real life vector compression applications [72].

Distributed Lossy Compression

In an uplink distributed MIMO C-RAN system, the received signals at each receiver must be individually compressed and forwarded to the CP. These received signals are inherently statistically dependent, and a good compression scheme should account for this to reduce redundant signal features for efficient fronthaul use. Under Gaussian vector compression, this dependency can be partially accounted for by jointly choosing the quantization noise covariance matrices to optimise overall MIMO performance.

However, point-to-point compression schemes are generally sub-optimal in the multi-receiver setting, with so-called 'distributed source coding' schemes having the potential for improved performance. These schemes have their roots in the landmark work on distributed lossless compression by Slepian & Wolf [199], and subsequent adaptation for lossy compression by

Wyner & Ziv [232], characterised by their use of *joint* decompression for recovering the set of compressed signals.

In the canonical Wyner-Ziv-type compression scheme, quantization is first applied at each source to produce lossy signal representations. The individual encoders then map the finite number of quantized signal states to a *reduced* number of codewords, or 'bins', using a many-to-one mapping, therefore reducing the encoding rate. By jointly considering the codewords from all sources, the decoder can exploit signal correlations to accurately determine the original quantized signal state for each codeword and recover the compressed signals.

However, its performance gains come at the expense of increased computational complexity at the decoder, which may be prohibitive in large distributed systems. Distributed source coding has previously been applied in compression for video applications and sensor networks, and has been considered for use in C-RAN systems.

4.2.3 Signal Compression for Uplink MIMO C-RAN

A good fronthaul compression scheme maximises MIMO performance under given fronthaul capacity constraints. To evaluate the performance of such schemes it is useful to compare them to theoretical performance bounds from network information theory. The cut-set bound states that the 'rate of information flow across any boundary is less than the mutual information between the inputs on one side of the boundary and the outputs on the other side, conditioned on the inputs on the other side' [47]. A useful upper bound follows from this – the total sum capacity of the MIMO C-RAN system cannot exceed either the sum capacity of the uncompressed wireless links, or the total fronthaul capacity, as shown in 4.4.

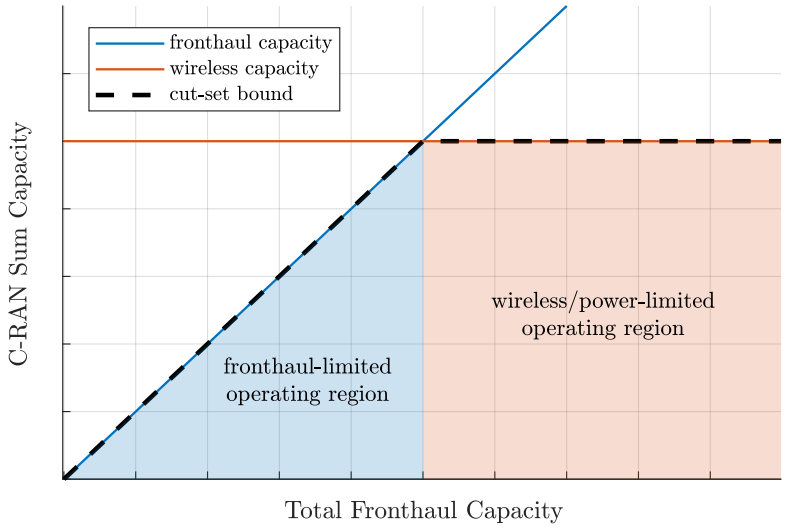

Figure 4.4: Cut-set upper bound on sum capacity of C-RAN network with limited fronthaul capacity.

At low fronthaul rates the system sum capacity is fundamentally limited by the amount of fronthaul available. Within this *fronthaul-limited* operating region, efficient signal compression is vital to ensure good system capacity is achieved. At higher fronthaul capacities, the

fundamental limit is instead the capacity of the wireless links – or the transmit power of the users. The fronthaul efficiency naturally drops off in this operating region, and hence efficient compression is necessary to ensure good fronthaul utilisation. In many scenarios the cut-set bound cannot be achieved by compress and forward schemes [53], but the bound serves as a useful indicator of performance.

For a distributed MIMO system with L receivers connected to the CP by fronthaul links with capacity $\mathcal{R}$, the sum capacity cut-set bound is

$$\mathcal{C}_{\text{SUM}} = \mathcal{I}\big(\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_L; \mathbf{x}\big) \tag{4.24}$$

$$\leq \min\Big(\mathcal{I}\big(\mathbf{y}_1, \ldots, \mathbf{y}_L; \mathbf{x}\big), \mathcal{R}L\Big). \tag{4.25}$$

Of course, in a cellular MU-MIMO system other metrics beyond sum capacity, such as per-user and outage capacities, are also of importance. Moreover, the fronthaul efficiency of a compression scheme is not its only important feature – cellular systems operate with constraints on delay, in mobile environments where the channels may change rapidly, and therefore both computational complexity and the signalling overheads associated with a scheme are also important.

A variety of different approaches have been taken to solving this problem, ranging from practical quantization-based schemes to more complex distributed coding schemes that establish achievable bounds on performance. This section outlines the most significant of these schemes. Compression strategies based on dimension reduction are the subject of the next chapter, to which discussion of them is deferred.

Low Resolution Sampling

The most direct way of reducing uplink fronthaul traffic is to reduce the amount of sampled data produced. A considerable amount of research has been devoted to the use of reduced resolution and single-bit ADCs for sampling the received signals in MIMO systems. This work has mainly focused on the potential reduction in hardware cost and improved energy efficiency, but the potential benefits for reducing fronthaul load have also been noted. For example, in [202] is is shown that reducing the ADC resolution to 4-6 bits per I or Q sample (8-12 bits per antenna) in a massive MIMO-OFDM systems incurs negligible performance loss, whilst reducing the sampled data by around half compared to conventional high-resolution sampling (10+ bits per I or Q). The work in [164] considers the use of analogue beamforming and single-bit ADCs for reducing the fronthaul traffic in a cell-free MIMO C-RAN system.

Explicit optimisation of the sampling resolution for fronthaul capacity constrained systems is carried out in [163], where a limited number of bits are optimally allocated to the ADCs at the receive antennas. Improved fronthaul efficiencies are demonstrated, but the analysis is limited to narrowband MIMO systems, and it is unclear whether such an architecture could in practice be implemented in hardware. The optimal number of antennas under equal resolution ADCs with a fixed total number of sampling bits is considered in [99], concluding that this varies according to SNR.

Overall, whilst offering some gains, low resolution sampling is inflexible and cannot properly

exploit the inherent temporal & spatial correlations in the received signals. Most work on fronthaul-constrained C-RAN MIMO therefore considers signal compression as a secondary stage that is implemented digitally after sampling. The performance degradation that occurs due to digitisation is then typically ignored when designing fronthaul compression schemes (implicitly assuming ADC quantization noise is small compared to receiver noise).

Functional Splits & Signal Quantization Methods

Improved compression performance can be achieved by performing some functions at the remote receiver before transferring bits over fronthaul. In [182] filtering and downsampling is used to remove temporal redundancies due to oversampling, followed by digital quantization of the signal with adaptively chosen resolution. In [39], a classical principal component transform coding approach is used to identify correlations in the signal across both space and time, and reduce the signal dimensionality. Performing OFDM demodulation at the receiver is a simple way to remove redundancies due to oversampling and unused spectral resources [44]. This is a starting point for most of the proposed MIMO C-RAN fronthaul compression schemes, which then consider further compression of narrowband MIMO signals.

For single cell massive MIMO C-RAN systems, the large number of antennas at the receiver massively increases the amount of sampled data produced. In [95] & [32] it is noted that ZF/MMSE detection can be decomposed into a matched filtering stage, and an inversion stage. Since the matched filtering stage reduces the signal dimensionality from M to K, it is beneficial to perform this stage at the receiver to reduce the amount of data transferred across fronthaul, with the more computationally complex inversion stage performed at the CP.

The fronthaul reduction problem for distributed MIMO differs somewhat – the sampled antenna data being distributed between the different receivers. In [12] simple fixed-rate scalar quantization is used to reduce fronthaul load, applied to the subcarrier level received signal either before or after matched filtering, depending on the number of antennas per receiver and the number of users. In [132] it is shown that the use of ZF detection, with only a small number of bits and quantized channel estimates, is sufficient to outperform ideal unquantized MF-based processing – highlighting the benefit of performing full detection at the CP. In [121] bits are optimally allocated to the scalar quantizers on different subcarriers to maximise throughput.

The schemes above use a somewhat heuristic approach to reducing fronthaul load, but improved performance can be gained by approaching the problem from a signal compression theory perspective.

Point-to-Point Compression Schemes

In [248] optimal point-to-point fronthaul compression for distributed MIMO C-RAN with multi-antenna receivers under the Gaussian vector test channel compression model is considered. It is shown that maximising either sum capacity or weighted user rates subject to fronthaul capacity constraints involves solving a non-convex optimisation problem to find the optimal quantization noise covariance matrices. A stationary point to this problem is then shown to be found by solving alternating convex problems. Whilst this scheme effectively solves the point-to-point compression problem, it requires the use of computationally intensive numerical solvers and

must be recalculated for each coherence interval, making it impractical to implement in real systems. To address this, [249] shows that at high signal-to-quantization-noise ratio (SQNR), i.e. at higher fronthaul rates when the quantization noise is small compared to the desired signal, a simple uniform quantization noise level is asymptotically optimal, and can be found using a simple bisection method.

A more tractable suboptimal 'spatial compression & forward' approach for distributed multi-antenna MIMO C-RAN is investigated in [123]. This is effectively a transform coding method, where a linear transform is applied to the signal at each receiver, followed by fixed-rate scalar quantization. In their formulation, the linear transforms are calculated based on local CSI and fixed, with quantizer rate allocations and user transmit powers then optimised on a global level using a max-min SINR objective. A stationary point is also found by numerically solving a sequence of alternating convex optimisations, but on lower dimensional scalar bit and power optimisation variables, rather than matrix values. It is shown that when the network has an excess of antennas, the main benefit comes from the optimal allocation of quantization bits, and that with multiple distributed receivers the scheme outperforms a single massive MIMO receiver (with no fronthaul constraints). The benefits of deploying additional antennas at each remote receiver are also shown.

To reduce the hardware requirements when each receiver has a large number of antennas, a hybrid spatial compression scheme that uses both analogue and digital spatial filters is proposed in [118]. The digital filtering stage and rate allocation use instantaneous CSI and are formulated similarly to [123], whilst the analogue stage performs a dimension reduction and is adapted on a slower timescale to match the channel statistics. Simulations show that the performance of the proposed scheme is close to that of the fully-digital scheme, but benefits from reduced complexity.

Wyner-Ziv Compression Schemes

The best compression performance is achieved using Wyner-Ziv type distributed source coding with joint signal decompression, and has received attention from the information theory community for both gaining theoretical insights and for the design of compression schemes.

Amongst early work, the theoretical sub-optimality of using compress and forward with Gaussian signalling in fronthaul-constrained networks is demonstrated in [183]. Nonetheless, the compress and forward strategy has attracted attention, due to its tractability and consistency with practical transmission schemes.

When the input signal is constrained to be Gaussian, [247] shows that the Gaussian vector Wyner-Ziv compression represents the optimal compression. However, it is noted that the number of constraints involved in determining the quantization noise covariances grows exponentially with the number of fronthaul links, and hence optimal compression is impractical for all but the smallest networks. To reduce the number constraints for the Wyner-Ziv scheme, [249] considers a system in which compressed signals from receiver are decompressed sequentially, using the previously decompressed signals as side information. As with the point-to-point case, a uniform quantization noise level is shown to be asymptotically optimal at high SQNR.

In [49], an aggregate fronthaul capacity constraint is instead considered (i.e. shared fron-

thaul), and optimal compression shown to be achieved by transform coding using the conditional Karhunen Loeve transform (KLT), and allocating rates by waterfilling on the conditional eigenvalues. A similar scheme is implemented under individual fronthaul constraints with sequential decompression in [165], and then adapted to account for uncertainties in CSI.

4.3 Massive MIMO Uplink Signal Compression

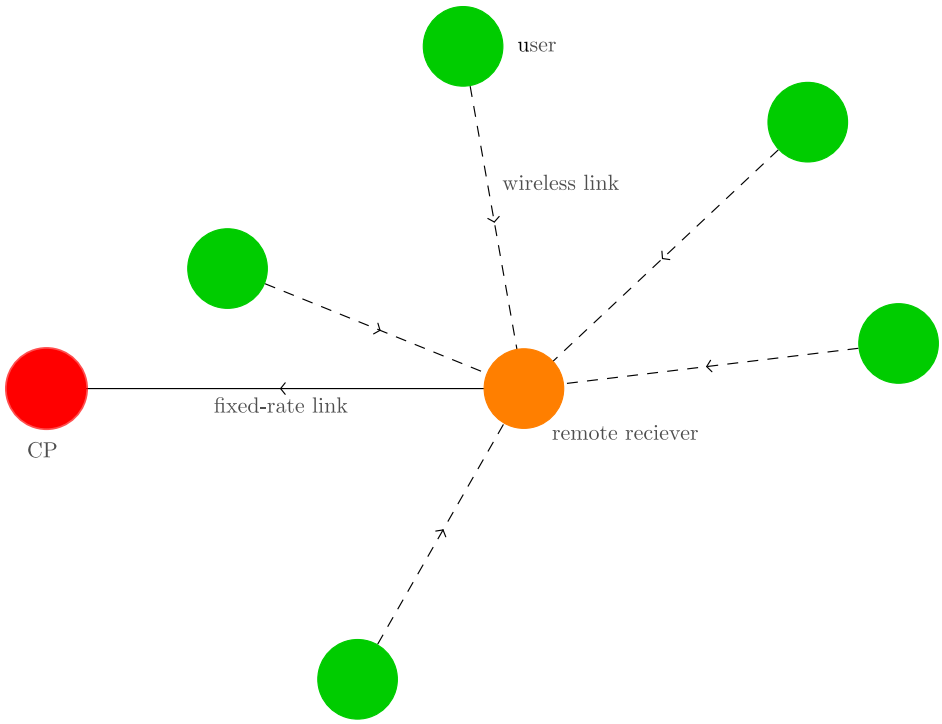

Figure 4.5: Single cell C-RAN uplink network topology

Whilst it is apparent that performing full MIMO detection at the receiver, i.e. a decode & forward strategy, would enable the cut-set bound to be fully achieved in the architecture in Figure 4.5, the use of compress & forward analysis has previously drawn interest within industry, e.g. [30], as a potential alternative in situations where for practical implementation reasons it is not desirable to implement full MIMO detection at the receiver. Furthermore, it serves as a useful starting point for analysing transform coding MIMO C-RAN systems, and to the analysis in Section 4.4 and the following chapter.

4.3.1 The Limits of Sample & Forward

Before considering more sophisticated compression schemes, it is insightful to provide a more theoretical motivation for their use, by establishing a sum capacity upper bound for the sample & forward case where the received signals are individually quantized using a set of M equal resolution scalar quantizers. This bound is based on compressing the signal at each antenna

independently using the optimal Gaussian scalar compression,

$$\tilde{y}_m = Q_m(y_m) \tag{4.26}$$

where the total available fronthaul bits, $\mathcal{R}$, are shared equally between all antennas, $r_m = \mathcal{R}/M$. Since the Gaussian scalar compression model provides a lower bound on quantization noise, it can be used to find an upper bound on the performance achieved by other sample & forward scalar compression schemes (e.g. conventional fixed-rate scalar quantization). The compressed signal can be represented by the Gaussian test channel,

$$\tilde{\mathbf{y}} = \mathbf{y} + \boldsymbol{\psi}, \tag{4.27}$$

where $\psi \sim \mathcal{CN}\big(0, \boldsymbol{\Psi}\big)$. The compression noise covariance matrix, $\boldsymbol{\Psi}$, is diagonal and from (4.9) has

$$\big[\boldsymbol{\Psi}\big]_{m,m} = \mathbb{E}\big[|y_m|^2\big] \frac{1}{2^{\mathcal{R}/M} - 1}, \tag{4.28}$$

giving

$$\boldsymbol{\Psi} = \big(\rho \mathbf{D}_{\mathrm{H}} + \mathbf{I}_M\big) \frac{1}{2^{\mathcal{R}/M} - 1} \tag{4.29}$$

where $\mathbf{D}_{\mathrm{H}}$ is a diagonal matrix containing the channel strengths for the M BS antennas, $\big[\mathbf{D}_{\mathrm{H}}\big]_{m,m} = \big[\mathbf{H}\mathbf{H}^{\dagger}\big]_{m,m}$.

The compressed sum capacity under Gaussian compression is given by

$$\begin{aligned} \mathcal{C}_{\mathrm{SUM}} &= \mathcal{I}\big(\tilde{\mathbf{y}}; \mathbf{x}\big) && (4.30)\\ &= \mathcal{H}\big(\mathbf{H}\mathbf{x} + \boldsymbol{\eta} + \boldsymbol{\psi}\big) - \mathcal{H}\big(\boldsymbol{\eta} + \boldsymbol{\psi}\big) && (4.31)\\ &= \log_2 \det\big(\mathbf{I}_K + \rho \mathbf{H}^{\dagger}\big(\boldsymbol{\Psi} + \mathbf{I}_M\big)^{-1}\mathbf{H}\big) && (4.32) \end{aligned}$$

An upper bound for this sum capacity,

$$\mathcal{C}_{\mathrm{SUM}} < \mathcal{C}_{\mathrm{SF}}^{\mathrm{UB}}, \tag{4.33}$$

can be found using three steps (see Appendix 2.1):

1. Upper bounding the capacity using

$$\log_2 \det\big(\mathbf{I}_K + \mathbf{A}\big) \leq \log_2 \det\big(\mathbf{A}\big) + \log_2(e)\, \mathrm{Tr}\big(\mathbf{A}^{-1}\big) \tag{4.34}$$

2. Lower bounding the quantization noise covariance as

$$\boldsymbol{\Psi} < \big(\rho \mathbf{D}_{\mathrm{H}} + \mathbf{I}_M\big) 2^{-\mathcal{R}/M} \tag{4.35}$$

3. Taking the limit as $\rho \to \infty$.

This results in

$$\mathcal{C}_{\mathrm{SF}}^{\mathrm{UB}} = \frac{\mathcal{R}K}{M} + \log_2 \det\left(\mathbf{H}^\dagger \mathbf{D}_{\mathrm{H}}^{-1} \mathbf{H}\right) + 2^{-\mathcal{R}/M} \log_2(e)\, \mathrm{Tr}\left(\left(\mathbf{H}^\dagger \mathbf{D}_{\mathrm{H}}^{-1} \mathbf{H}\right)^{-1}\right) \tag{4.36}$$

$$\approx \frac{\mathcal{R}K}{M} + \epsilon, \tag{4.37}$$

and is tight at high SNR. The third term in (4.36) decays rapidly with $\mathcal{R}$, and therefore the sum capacity bound increases approximately linearly with the fronthaul capacity, $\mathcal{R}$, at gradient K/M. For a massive MIMO system, where $M/K \gg 1$, the capacity therefore increases slowly with the available fronthaul. This is illustrated in Figure 4.6 for the i.i.d Rayleigh channel with $K = 8$ users.

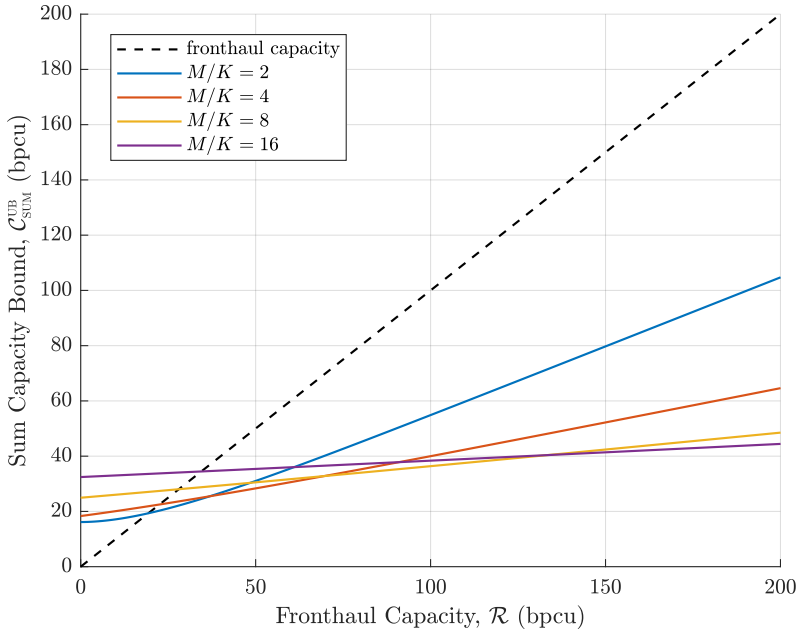

Figure 4.6: Sum capacity upper bound under sample & forward for different antenna ratios, i.i.d Rayleigh fading, $K = 8$.

From the bound it can be concluded that in massive MIMO systems sampling/quantizing the signals at each antenna and forwarding them over fronthaul results in poor fronthaul efficiency[3], and a system that operates far from the cut-set bound. This is illustrated in Figure 4.7 for i.i.d Rayleigh fading with $K = 8$ and $M = 32$ (also showing that the bound becomes tight at high SNR).

[3]This analysis assumes sampling of narrowband MIMO signals. However, a similar bound is easy to demonstrate for the case of OFDM signals with time domain sampling. In this case the matrix of channel strengths, $\mathbf{D}_{\mathrm{H}}$, is replaced by a matrix containing the average received signal power for each antenna.

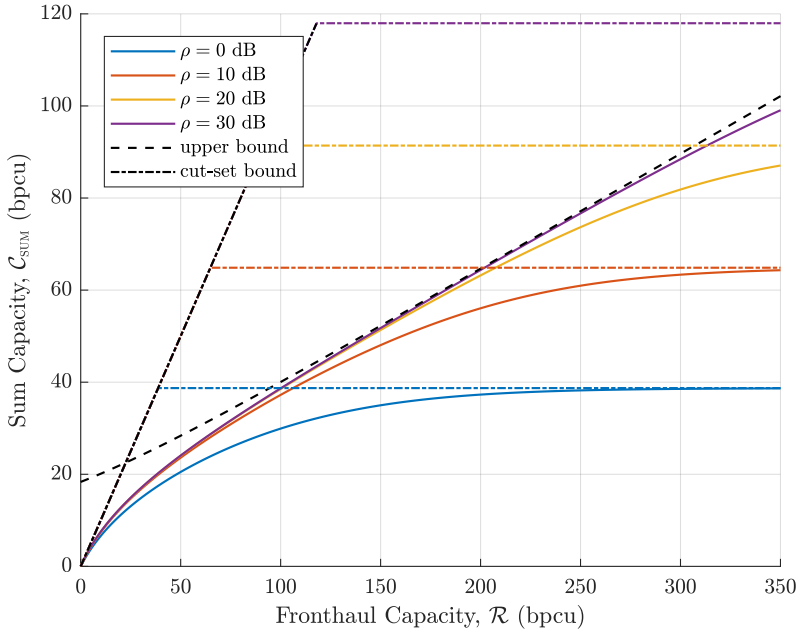

Figure 4.7: Sum capacity under sample & forward for different SNRs, i.i.d Rayleigh fading, $K = 8$, $M = 32$.

4.3.2 Transform Coding for the Massive MIMO Uplink

Transform coding is a widely used signal compression method, attractive for both its simplicity and performance. Its idea is to transform a set of correlated variables to a new signal basis, in which scalar compression of the transformed variables is more efficient [72]. This is best achieved when the signal exhibits *sparsity* – the underlying dimensionality of the signal is less than the original number of measurements, and a new signal basis can therefore be found in which the signal information is contained in a reduced number of variables.

The uplink signals of massive MIMO systems are characterised by their sparsity, having M measurements of K user symbols, where $M \gg K$. The previous section has shown that compressing all M signals results in poor fronthaul efficiency. Transform coding is therefore an intuitively good strategy for compressing uplink massive MIMO signals.

Transform coding has previously been proposed for single receiver massive MIMO signal compression in [39] & [161], but was applied in a fairly heuristic manner to the time domain samples. Here, it is proposed to apply transform coding to the multi-antenna signal on a subcarrier level[4], using a linear transform (or spatial filter), $\mathbf{F} \in \mathbb{C}^{M \times K}$, to produce a K-dimensional signal, $\mathbf{z} \in \mathbb{C}^{K}$,

$$\mathbf{z} = \mathbf{F}^{\dagger}\mathbf{y}. \tag{4.38}$$

[4]This follows since the signals on each subcarrier are generally independent in OFDM. The FFT can in fact be considered as part of the transform coding itself; optimal transform coding applied across space and time would incorporate the FFT to perform time domain decorrelation, before applying decorrelating in the spatial domain.

This is then compressed using a set of K Gaussian scalar compressors,

$$\tilde{z}_i = Q_i(z_i), \tag{4.39}$$

producing the compressed signal $\tilde{\mathbf{z}} \in \mathbb{C}^K$,

$$\tilde{\mathbf{z}} = \mathbf{z} + \boldsymbol{\delta}, \tag{4.40}$$

where $\boldsymbol{\delta} \sim \mathcal{CN}(0, \boldsymbol{\Phi})$ with $\boldsymbol{\Phi} = \mathrm{diag}(\phi_i)$ a diagonal matrix. MIMO detection may then be performed directly on $\tilde{\mathbf{z}}$.

The covariance of the received signal is

$$\mathbb{E}\big[\mathbf{y}\mathbf{y}^\dagger\big] = \rho\mathbf{H}\mathbf{H}^\dagger + \mathbf{I}_M \tag{4.41}$$

$$= \rho\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\dagger + \mathbf{I}_M, \tag{4.42}$$

where $\mathbf{U} \in \mathbb{C}^{M\times M}$ are the eigenvectors and diagonal $\boldsymbol{\Lambda} \in \mathbb{C}^{M\times M}$ the ordered eigenvalues, λ_i, as in (2.84). Since only K of these eigenvalues are non-zero, the user data signal lies entirely within the subspace spanned by the first K eigenvectors, $\mathbf{U}_{1:K} \in \mathbb{C}^{M\times K}$ (the other subspace containing only noise). If the columns of $\mathbf{F}$, $\mathbf{f}_i$, span this subspace then the transform is information-lossless,

$$\mathcal{I}(\mathbf{z}; \mathbf{x}) = \mathcal{I}(\mathbf{y}; \mathbf{x}). \tag{4.43}$$

If $\mathbf{F}$ is constrained to be semi-orthogonal,

$$\mathbf{F}^\dagger\mathbf{F} = \mathbf{I}_K, \tag{4.44}$$

then the transformed signal can be written

$$\mathbf{z} = \mathbf{F}^\dagger\mathbf{H}\mathbf{x} + \boldsymbol{\eta}_{\mathrm{F}}, \tag{4.45}$$

where $\boldsymbol{\eta}_{\mathrm{F}} \sim \mathcal{CN}\big(0, \mathbf{I}_K\big)$ is uncorrelated receiver noise.

With average bit rate r_i at scalar compressor i, the quantization noise variance is

$$\phi_i = \frac{\rho\mathbf{f}_i^\dagger\mathbf{H}\mathbf{H}^\dagger\mathbf{f}_i + 1}{2^{r_i} - 1}, \tag{4.46}$$

where the rates are allocated subject to the total fronthaul constraint,

$$\sum_{i=1}^{K} r_i = \mathcal{R}. \tag{4.47}$$

The massive MIMO transform coding scheme then effectively implements the Gaussian vector

test channel in Section 4.2.2, with

$$\mathbf{\Psi} = \begin{bmatrix} \mathbf{F} & \mathbf{U}_{K+1:M} \end{bmatrix} \begin{bmatrix} \phi_1 & & & \\ & \ddots & & \\ & & \phi_K & \\ & & & \boldsymbol{\infty} \end{bmatrix} \begin{bmatrix} \mathbf{F}^\dagger \\ \mathbf{U}_{K+1:M}^\dagger \end{bmatrix}, \tag{4.48}$$

where $\mathbf{U}_{K+1:M}$ are the eigenvectors corresponding to the noise subspace, the contents of which are discarded (modelled by infinite quantization noise).

Two transform coding schemes are now analysed. The first scheme known to maximise the sum capacity, whilst a second scheme, which uses a uniform quantization noise (UQN), is shown to be near-optimal and useful for providing analytical insights.

Sum Capacity Maximising Transform Coding

The mutual information maximising compression scheme – amongst *all* potential compression schemes – for a Gaussian vector source is shown in [226] to use transform coding. However, the application of this optimal transform coding scheme to massive MIMO systems has not previously been studied in detail. A simplified derivation (that does not prove global optimality) of this scheme is now provided, along with some observations specific to the C-RAN fronthaul problem.

Intuitively, the best signal basis for scalar compression has components of $\mathbf{z}$ that are statistically independent. Using the first K eigenvectors as a transform, $\mathbf{F} = \mathbf{U}_{1:K}$, produces a signal with diagonal covariance

$$\mathbb{E}\big[\mathbf{z}\mathbf{z}^\dagger\big] = \mathbf{\Lambda}_{1:K} + \mathbf{I}_K. \tag{4.49}$$

This transform is known as the Karhunen-Loeve transform (KLT) and is well known to be optimal in rate-distortion compression, finding wide use in transform coding applications [72]. The optimality of the KLT for transform coding under a sum capacity measure is shown in [226]. Since the outputs of the KLT are independent, the sum capacity can be expressed in the familiar eigenchannel form of (2.85),

$$\mathcal{C}_{\text{SUM}} = \sum_{i=1}^{K} \mathcal{I}(\tilde{z}_i; \mathbf{x}) \tag{4.50}$$

$$= \sum_{i=1}^{K} \log_2 \Big(1 + \frac{\rho\lambda_i}{1+\phi_i}\Big). \tag{4.51}$$

where each channel is degraded by the quantization noise,

$$\phi_i = \frac{\rho\lambda_i + 1}{2^{r_i} - 1}, \tag{4.52}$$

leading to

$$\mathcal{C}_{\text{SUM}} = \sum_{i=1}^{K} \log_2 \Big(1 + (1 - 2^{-r_i}) \frac{\rho\lambda_i}{\rho\lambda_i 2^{-r_i} + 1}\Big) \tag{4.53}$$

Asymptotically, at high SNR

$$\lim_{\rho\lambda_i\to\infty} \sum_{i=1}^{K} \log_2\left(1+(1-2^{-r_i})\frac{\rho\lambda_i}{\rho\lambda_i 2^{-r_i}+1}\right) = \sum_{i=1}^{K} r_i = \mathcal{R}, \tag{4.54}$$

whilst as the scalar compression rates increase,

$$\lim_{r_i\to\infty} \sum_{i=1}^{K} \log_2\left(1+(1-2^{-r_i})\frac{\rho\lambda_i}{\rho\lambda_i 2^{-r_i}+1}\right) = \sum_{i=1}^{K} \log_2\left(1+\rho\lambda_i\right). \tag{4.55}$$

These two limiting cases correspond to the cut-set bound – it is asymptotically achieved by transform coding with the KLT at high SNR, under *any* rate allocation. This is because receiver noise decreases relative to quantization noise as SNR increases, so that at high SNR quantization noise becomes the only factor limiting performance. Figure 4.8 compares sum capacity to fronthaul capacity for different SNRs, where all quantities are normalised by dividing by the sum capacity of the respective uncompressed MIMO channel.

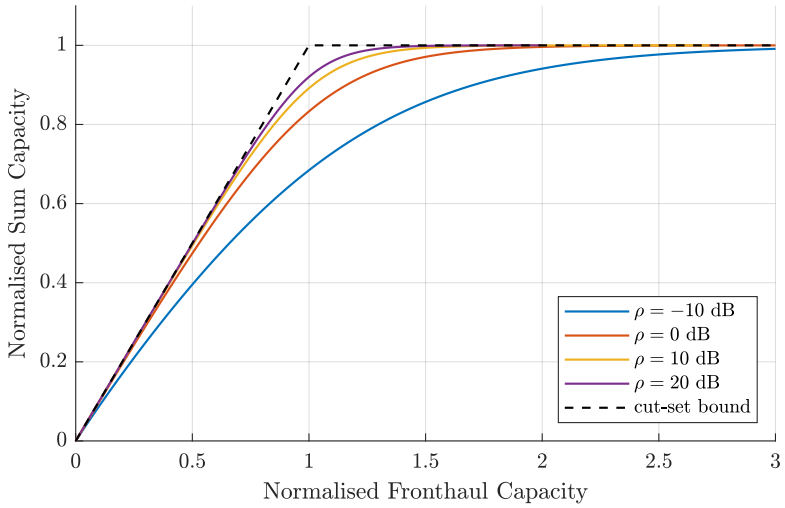

Figure 4.8: Normalised sum capacity of massive MIMO system under transform coding with varying SNR, i.i.d Rayleigh channel, $K = 8$, $M = 64$.

However, in the fronthaul-limited region, increasing the SNR to move towards the cut-set bound brings diminishing returns, and decreases in energy efficiency. At lower SNRs both receiver noise and quantization noise degrade performance, and appropriate rate allocation is required to maximise capacity. The optimal rate allocation under the fronthaul constraint can be shown using Lagrange multipliers to be the simple waterfilling solution,

$$r_i = \left[\frac{\mathcal{R}}{|\mathcal{S}|} + \log_2(\lambda_i) - \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \log_2(\lambda_j)\right]^+, \tag{4.56}$$

where $[a]^+ = \max(a, 0)$ and $\mathcal{S}$ is the set of i for which $r_i > 0$[5].

[5]This is found by starting with $|\mathcal{S}| = K$, allocating rates as in (4.56) and iteratively removing the indices corresponding to the smallest remaining eigenvalues from $\mathcal{S}$ until $\sum r_i = \mathcal{R}$

At low fronthaul rates, only a subset of the eigenchannels are compressed, an effect that is more pronounced when there is a large channel eigenvalue spread. This results in a MU-MIMO uplink that is degraded (rank deficient) at low fronthaul capacities. On the other hand, at higher fronthaul rates (when all eigenchannels are compressed) the allocated rates are

$$r_i = \frac{\mathcal{R}}{K} + \log_2(\lambda_i) - \frac{1}{K}\sum_{j=1}^{K}\log_2(\lambda_j), \tag{4.57}$$

resulting in quantization noise

$$\phi_i = \frac{\rho\lambda_i + 1}{\frac{\lambda_i}{\overline{\lambda}}2^{\mathcal{R}/K} - 1}, \tag{4.58}$$

where $\overline{\lambda} = \left(\prod_{i=1}^{K}\lambda_i\right)^{1/t}$ is the geometric mean of the eigenvalues. When the fronthaul capacity is sufficiently large that $\lambda_i/\overline{\lambda}2^{\mathcal{R}/K} \gg 1$, and at high SNR such that $\rho\lambda_i \gg 1$, this simplifies to[6]

$$\phi_i \approx \rho\overline{\lambda}2^{-\mathcal{R}/K}, \tag{4.59}$$

i.e. the quantization noise is approximately uniform across all eigenchannels.

In contrast to the sample & forward method, for a given SNR, ρ, fronthaul efficiency increases as the number of antennas increases, as shown in Figure 4.9. This is because as the number of antennas increases the magnitude of the eigenvalues increases[7], boosting the SNR, $\rho\lambda_i$, of each eigenchannel. However, this increase in sum capacity is limited by the fronthaul capacity in the fronthaul-limited region.

[6]This is actually a lower bound on ϕ_i that becomes tight as ρ and $\mathcal{R}$ increase, see Figure 4.10.
[7]See (4.71).

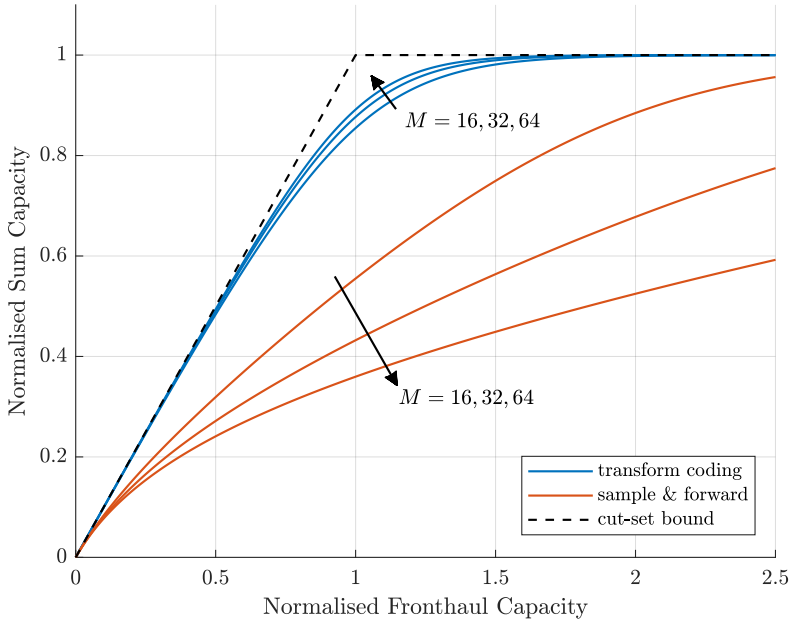

Figure 4.9: Comparison of sum capacity of massive MIMO with transform coding and sample & forward compression with varying number of antennas, M, i.i.d Rayleigh channel, $K = 8$, $\rho = 10$ dB.

4.3.3 Transform Coding with a Uniform Quantization Noise Level

Whilst the sum capacity maximising scheme is known, a uniform quantization noise (UQN) level,

$$\phi_i = \Delta, \quad \forall i \tag{4.60}$$

is attractive from an analysis perspective, since it is modelled by a simple uniform increasing in the noise power the MIMO system experiences – the sum capacity under transform coding with a UQN is simply

$$\mathcal{C}_{\text{SUM}} = \log_2 \det \left(\mathbf{I}_K + \frac{\rho}{1+\Delta} \mathbf{H}^\dagger \mathbf{F} \mathbf{F}^\dagger \mathbf{H} \right) \tag{4.61}$$

$$= \log_2 \det \left(\mathbf{I}_K + \frac{\rho}{1+\Delta} \mathbf{H}^\dagger \mathbf{H} \right) \tag{4.62}$$

$$= \sum_{i=1}^{t} \log_2 \left(1 + \frac{\rho \lambda_i}{1+\Delta} \right). \tag{4.63}$$

The allocated compression rates are

$$r_i = \log_2 \left(1 + \frac{\rho \mathbf{f}_i^\dagger \mathbf{H} \mathbf{H}^\dagger \mathbf{f}_i + 1}{\Delta} \right), \tag{4.64}$$

where the minimum Δ can be found that satisfies the fronthaul constraint, $\sum r_i = \mathcal{R}$, can be found using, for example, a simple bisection algorithm as shown in Algorithm 2.

Algorithm 2 Bisection algorithm for finding UQN level, Δ

inputs: $\mathbf{H}, \mathbf{F}, \mathcal{R}$
initialise $\Delta_{\min}$, $\Delta_{\max}$
while $\left|\sum_i r_i - \mathcal{R}\right| \geq \epsilon$ **do**
  $\Delta = \dfrac{\Delta_{\min} + \Delta_{\max}}{2}$
  $r_i = \log_2\left(1 + \frac{\rho \mathbf{f}_i^\dagger \mathbf{H}\mathbf{H}^\dagger \mathbf{f}_i + 1}{\Delta}\right)$
  if $\sum_i r_i > \mathcal{R}$ **then**
    $\Delta_{\min} = \Delta$
  else
    $\Delta_{\max} = \Delta$
  end if
end while
outputs: Δ, r_i

At all but low fronthaul rates

$$r_i \approx \log_2\left(\frac{\rho \mathbf{f}_i^\dagger \mathbf{H}\mathbf{H}^\dagger \mathbf{f}_i + 1}{\Delta}\right), \tag{4.65}$$

and the UQN level is approximately given by,

$$\Delta \approx \left(\prod_{i=1}^{K} \left(\rho \mathbf{f}_i^\dagger \mathbf{H}\mathbf{H}^\dagger \mathbf{f}_i + 1\right)\right)^{1/K} 2^{-\mathcal{R}/K}. \tag{4.66}$$

Thus whilst the sum capacity can be expressed solely in terms of $\mathbf{H}$, ρ and Δ, it also depends on the transform used, $\mathbf{F}$, through the dependence of the UQN compression noise level in (4.66). Applying Hadamard's inequality[8] the KLT minimises Δ in (4.66), giving

$$\Delta \approx \left(\prod_{i=1}^{K} \left(\rho \lambda_i + 1\right)\right)^{1/K} 2^{-\mathcal{R}/K}. \tag{4.67}$$

This is approximately equal to that achieved by the optimal rate allocation in (4.59) when all $\rho\lambda_i \gg 1$, and hence the UQN rate allocation also approximately maximises sum capacity. This is shown in Figure 4.10, which compares the eigenchannel quantization noise with KLT transform under waterfilling and UQN rate allocations, for the i.i.d Rayleigh fading channel with $K = 8$, $M = 64$, $\rho = 10$ dB. Under waterfilling at low rates some of the eigenchannels are not compressed ($r_i = 0$), and therefore have infinite quantization noise variance. At higher rates, UQN and waterfilling give the same quantization noise and the approximation in (4.59) is tight.

Figure 4.11 shows the sum capacity achieved under both waterfilling and UQN rate al-

[8]Hadamard's inequality [91] for positive-semidefinite matrices states that the product of the diagonal entries of a matrix is lower bounded by the determinant of the matrix. Here, $\prod_i \left(\rho \mathbf{f}_i^\dagger \mathbf{H}\mathbf{H}^\dagger \mathbf{f}_i + 1\right) \geq \prod_i \left(\rho\lambda_i + 1\right)$, with equality when $\mathbf{F} = \mathbf{U}_{1:K}$.

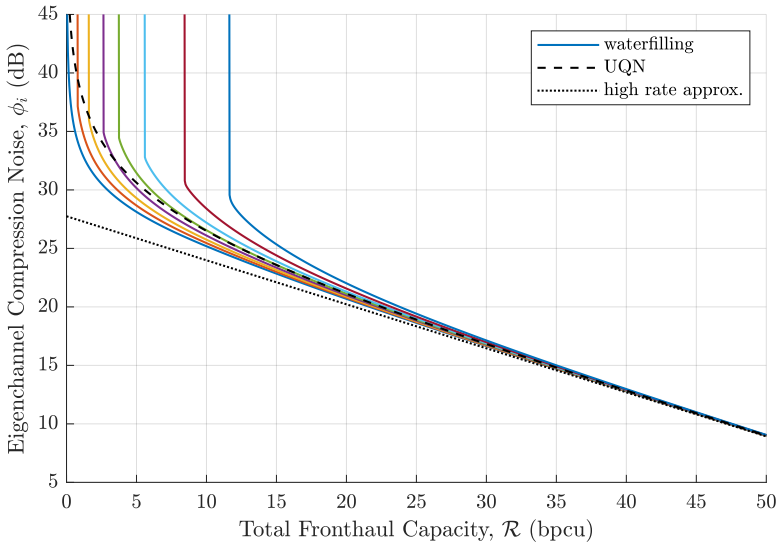

Figure 4.10: Scalar quantization noise variances under waterfilling and UQN rate allocations, i.i.d Rayleigh fading channel, $K = 8$, $M = 64$, $\rho = 10$ dB. Coloured lines show the quantization noises on different eigenchannels.

locations, for a range of SNRs. At all SNRs, the performance difference between UQN and waterfilling is negligible, and the sum capacity comes close to the cut-set bound. For instance with $\rho = 0$ dB, a fronthaul capacity of 40 bpcu achieves sum capacity of 35 bpcu, a fronthaul efficiency of 87 %, and with $\rho = 10$ dB a fronthaul capacity of 60 bpcu can achieve sum capacity of 57 bpcu, a fronthaul efficiency of 95 %. Transform coding significantly outperforms the sample & forward scheme at all fronthaul rates and SNRs.

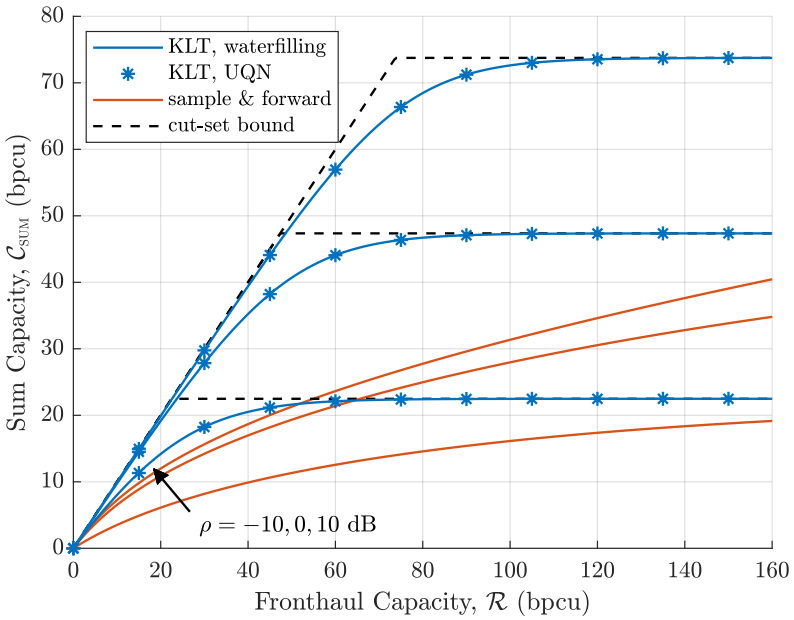

Figure 4.11: Sum capacity under transform coding with KLT transform and waterfilling & UQN rate allocations, i.i.d Rayleigh fading channel, $K = 8$, $M = 64$.

4.3.4 Massive MIMO with Limited Fronthaul

The work so far has established that transform coding can be an efficient fronthaul compression technique for massive MIMO C-RAN systems. However, as Figure 4.9 shows, in the fronthaul-limited region, the sum capacity improvements available from increasing the number of BS antennas is fundamentally limited by the fronthaul capacity. Thus a key question remains: *if an uplink MIMO C-RAN system has limited fronthaul capacity, what, if any, are the benefits of deploying a large number of BS antennas?*

A simple analysis of massive MIMO under transform coding with UQN rate allocation provides some answers to this question. The key insight used is that under the UQN rate allocation the effect of compression is simply to uniformly increase the effective noise level seen in the MIMO system. Since this increase in noise level does not modify the MIMO channel in any way, a number of the benefits of favourable propagation, outlined in Section 2.5.1, are preserved under UQN fronthaul compression. Specifically, as the number of antennas is increased:

- Linear processing achieves the available sum capacity at any fronthaul rate and SNR.
- The user transmit power required to achieve a given service level reduces.
- The variations in per-user capacity due to fast fading disappear.

As a consequence, when operating in the fronthaul-limited region with practical linear processing both the user mean and outage capacities are significantly improved by increasing the number of BS antennas. This is shown in Figure 4.12 with $K = 8$, fixed fronthaul capacity $\mathcal{R} = 24$ bpcu and MMSE detection for mean and 10% outage capacities.

The underlying claims are now explored in more detail.

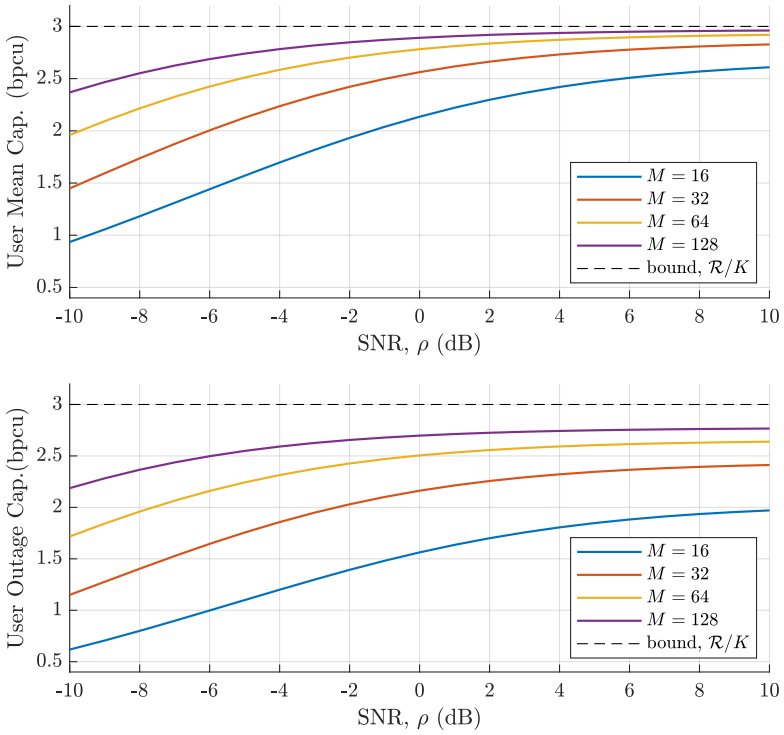

Figure 4.12: Mean and outage user capacities with varying number of BS antennas, M, fixed fronthaul capacity $\mathcal{R} = 24$ bpcu, i.i.d Rayleigh fading, $K = 8$.

Optimality of Linear Processing

In the previous section the achievable MIMO sum capacity under transform coding was studied, where it was assumed that optimal non-linear detection (MMSE-SIC) was used. However, non-linear methods do not scale well to MIMO systems with large numbers of users, and hence linear methods are desirable in practice.

Since the only effect of compression under the UQN rate allocation is to increase the system noise level, the optimality of linear processing, at any SNR and fronthaul rate, follows directly from the reasoning provided in Chapter 2. This linear detection can be performed directly on $\tilde{\mathbf{z}}$,

$$\hat{\mathbf{x}} = \mathbf{W}\tilde{\mathbf{z}}, \tag{4.68}$$

using, for example, MMSE detection with

$$\mathbf{W}^{(\mathrm{MMSE})} = \frac{1}{1+\Delta}\left(\mathbf{H}^{\dagger}\mathbf{H} + \frac{1+\Delta}{\rho}\mathbf{I}_K\right)^{-1}\mathbf{H}^{\dagger}\mathbf{F}. \tag{4.69}$$

Using (2.106), the per-user capacity under MMSE detection is given by

$$\mathcal{C}_k^{(\mathrm{MMSE})} = -\log_2\left(\left[\left(\mathbf{I}_K + \frac{\rho}{1+\Delta}\mathbf{H}^{\dagger}\mathbf{H}\right)^{-1}\right]_{k,k}\right). \tag{4.70}$$

Figure 4.13 compares the sum capacities achievable under optimal detection and MMSE detection for different numbers of BS antennas, at both high and low SNR. In the massive MIMO

regime ($M \geq 32$) the performance penalty from using linear detection is negligible at all fronthaul rates.

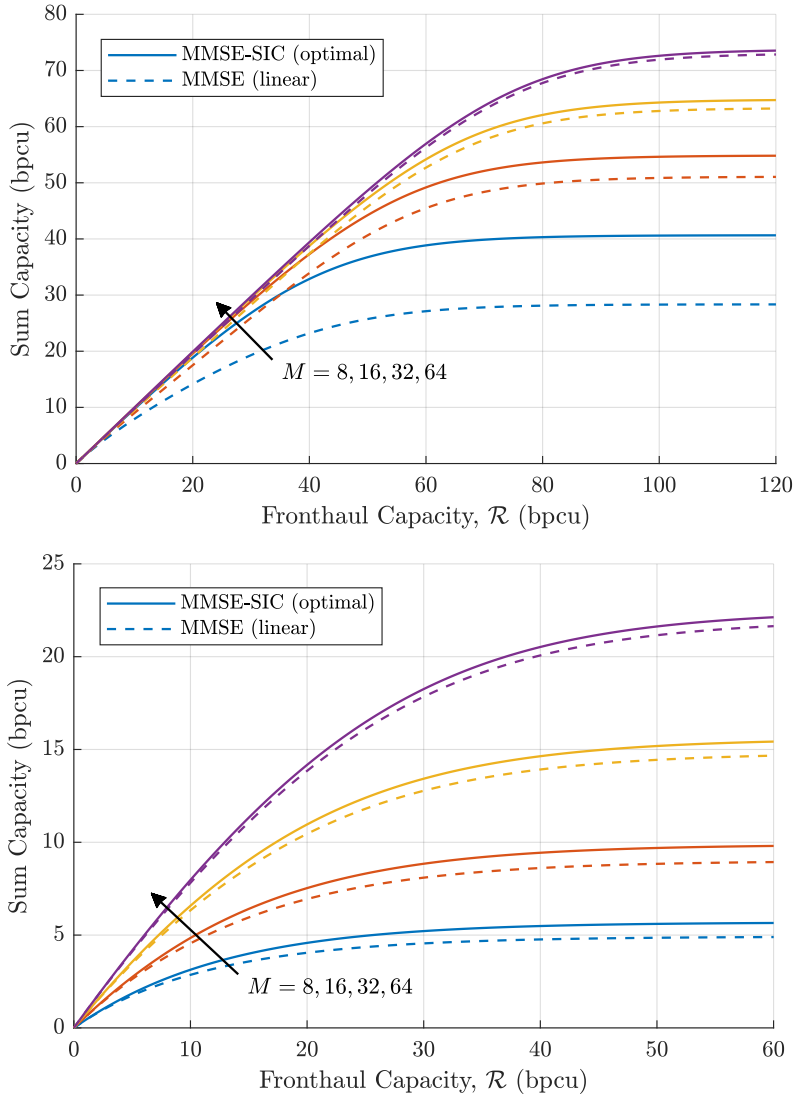

Figure 4.13: Sum capacity under non-linear and linear detection for different numbers of antennas, i.i.d Rayleigh fading, $K = 8$. Top figure: $\rho = 10$ dB, bottom figure: $\rho = -10$ dB.

Transmit Power Reduces

When operating the MIMO system in the fronthaul-limited region the system capacity is fundamentally bounded by the fronthaul capacity, irrespective of the number of BS antennas. As (4.50) shows, the performance depends on the channel eigenvalues, λ_i, and SNR, ρ. Since the eigenchannel magnitudes scale with the number of receive antennas,

$$\sum_{i=1}^{K} \lambda_i = \|\mathbf{H}\|^2 \approx M \sum_{k=1}^{K} p_k \beta_k, \tag{4.71}$$

the user transmit power can be scaled $\sim 1/M$ whilst approximately maintaining a given level of performance. This is shown in Figure 4.14, where under a fixed fronthaul capacity doubling the number of BS antenna enables the user transmit power to be reduced by ~ 3 dB. This indicates that in the fronthaul-limited region deploying more antennas can be used to improve the energy efficiency of the network.

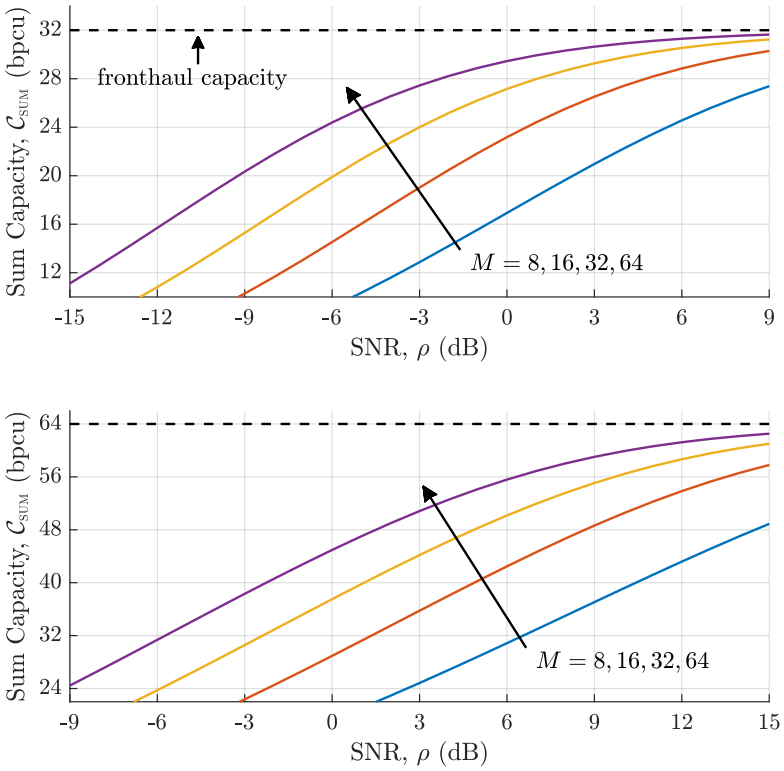

Figure 4.14: Sum capacity scaling with SNR for different numbers of antenna under a fixed fronthaul capacity, i.i.d Rayleigh fading, $K = 8$. Top figure: $\mathcal{R} = 32$ bpcu, bottom figure: $\mathcal{R} = 64$ bpcu.

Fast Fading Disappears

In the fronthaul limited region, the average user capacity cannot exceed

$$\mathbb{E}[\mathcal{C}_k] \leq \frac{\mathcal{R}}{K}, \tag{4.72}$$

but for a given channel realisation the individual user capacities may vary around this quantity. Since favourable propagation characteristics hold under UQN compression, channel hardening eliminates fast fading as M increases and, assuming operation close to the cut set bound, leads to

$$\mathcal{C}_k \to \frac{\mathcal{R}}{K}. \tag{4.73}$$

Thus, the users benefit from having near-deterministic individual capacities when many BS antennas are deployed. This is illustrated in Figure 4.15 for a fixed fronthaul rate of $\mathcal{R} = 24$

bpcu, where MMSE detection is used and SNR is backed off as $1/M$.

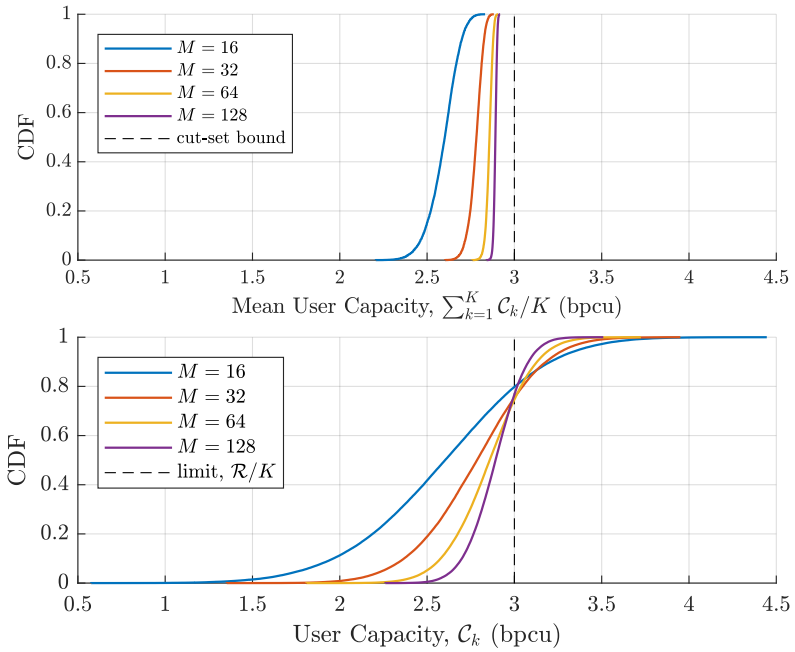

Figure 4.15: User capacity CDF for different numbers of antennas under a fixed fronthaul capacity $\mathcal{R} = 24$ bpcu, i.i.d Rayleigh fading channel, $K = 8$.

Correlated Channels

The above principles hold under any channel conditions that provide favourable propagation conditions. This is illustrated in Figure 4.16 for correlated Rayleigh fading with a uniform linear array and $K = 8$ users spaced at approximately equal angles over a $2\pi/3$ sector (one sector in a three sector cell), with equal pathloss. Linear processing improves as the number of antennas increases, but a larger number of antennas are required compared to the i.i.d case for it to become optimal, since favourable propagation occurs more slowly in this channel.

Similarly, the user mean and outage capacities improve as the number of BS antennas is increased, although the performance is not as good as in the i.i.d case, as shown in Figure 4.17.

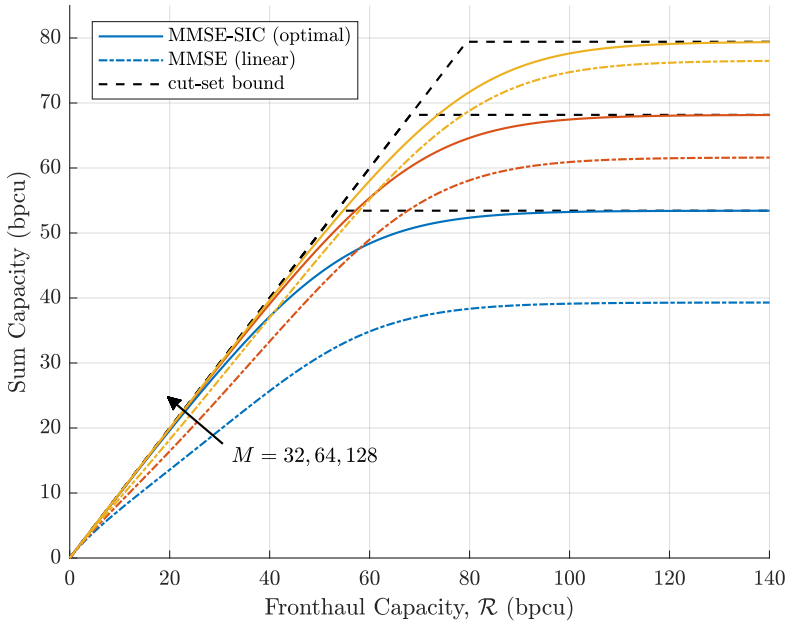

Figure 4.16: Sum capacity under non-linear and linear detection for different numbers of antennas, correlated Rayleigh fading, $\rho = 10$ dB, $K = 8$.

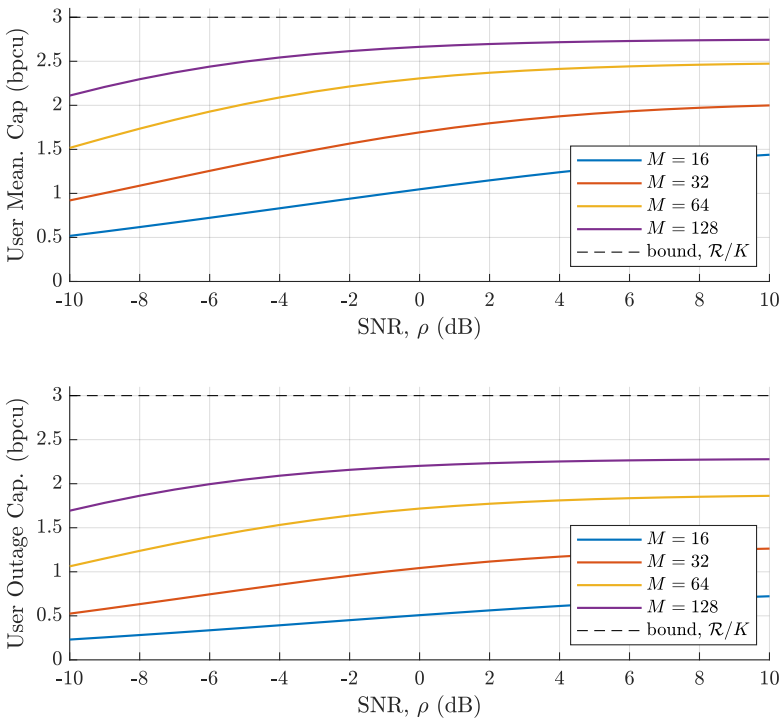

Figure 4.17: Mean and outage user capacities with varying number of BS antennas, M, fixed fronthaul capacity $\mathcal{R} = 24$ bpcu, correlated Rayleigh fading, $K = 8$.

4.3.5 Transform Coding with Imperfect CSI

The analysis provided so far assumes that the MIMO system has access to perfect CSI, when in practice this CSI must be estimated. Assuming that uplink pilots and MMSE channel estimation is used to estimate the channel matrix, the transform coding method can be adapted for the case of imperfect CSI using the additive channel estimation error model outlined in Section 2.4.2.

Transform Coding with UQN Rate Allocation

Under the additive channel error model, the received signal is

$$\mathbf{y} = \hat{\mathbf{H}}\mathbf{x} + \mathbf{E}\mathbf{x} + \boldsymbol{\eta} \tag{4.74}$$

$$= \hat{\mathbf{H}}\mathbf{x} + \boldsymbol{\nu}. \tag{4.75}$$

where $\hat{\mathbf{H}}$ is the estimated channel, $\mathbf{E}$ the channel estimation error, and $\boldsymbol{\nu}$ is equivalent noise with

$$\mathbb{E}_{\mathbf{E}}\left[\boldsymbol{\nu}\boldsymbol{\nu}^{\dagger}\right] = \boldsymbol{\Omega} = \mathbf{I}_M + \rho\sum_{k=1}^{K} p_k \mathbf{C}_k, \tag{4.76}$$

where $\mathbf{C}_k$ is the channel estimation error covariance for user k, and the expectation $\mathbb{E}_{\mathbf{E}}[.]$ is taken with respect to symbols, receiver noise and channel estimation error. A whitening transform can be used to whiten this equivalent noise[9]

$$\check{\mathbf{y}} = \boldsymbol{\Omega}^{-1/2}\mathbf{y} = \check{\mathbf{H}}\mathbf{x} + \check{\boldsymbol{\nu}}, \tag{4.77}$$

where $\check{\mathbf{H}} = \boldsymbol{\Omega}^{-1/2}\hat{\mathbf{H}}$ is the 'whitened' channel matrix. The decorrelating transform is then applied to the whitened signal,

$$\mathbf{z} = \mathbf{F}^{\dagger}\check{\mathbf{y}}. \tag{4.78}$$

The first K eigenvectors of the whitened channel matrix, $\mathbf{F} = \check{\mathbf{U}}_{1:K}$, decorrelate the whitened signal in the sense that

$$\mathbb{E}_{\mathbf{E}}\left[\mathbf{z}\mathbf{z}^{\dagger}\right] = \rho\check{\boldsymbol{\Lambda}}_{1:K} + \mathbf{I}_K \tag{4.79}$$

where $\check{\boldsymbol{\Lambda}} = \mathrm{diag}(\check{\lambda}_i)$ are the eigenvalues of $\check{\mathbf{H}}$. However, for a given channel estimate the signal components are not perfectly decorrelated, and the true instantaneous signal covariance

$$\mathbb{E}\left[\mathbf{z}\mathbf{z}^{\dagger}\right] = \rho\mathbf{F}^{\dagger}\boldsymbol{\Omega}^{-1/2}\mathbf{H}\mathbf{H}^{\dagger}\boldsymbol{\Omega}^{-1/2}\mathbf{F} + \mathbf{F}^{\dagger}\boldsymbol{\Omega}^{-1}\mathbf{F} \tag{4.80}$$

$$\neq \rho\check{\boldsymbol{\Lambda}}_{1:K} + \mathbf{I}_K \tag{4.81}$$

is unknown to the receiver, due to the imperfect CSI knowledge.

[9] This requires the singular value decomposition of an $M \times M$ matrix, which has complexity $\mathcal{O}(M^3)$. However, this can be reduced to $\mathcal{O}(K^3)$ using the fact that only the component lying in the K-dimensional signal subspace is required – for example, by taking a projection of $\mathbf{y}$ and $\boldsymbol{\Omega}$ into the channel subspace. Full details are omitted here for space.

Applying scalar compression to the elements of $\mathbf{z}$, gives

$$\tilde{\mathbf{z}} = \mathbf{F}^{\dagger}\check{\mathbf{H}}\mathbf{x} + \check{\boldsymbol{\nu}} + \boldsymbol{\delta}. \tag{4.82}$$

This compression model assumes that the instantaneous scalar variances $\sigma_i^2 = \mathbb{E}[|z_i|^2]$ are known. When a quantizer is designed for a source with differing statistics to the true source it is *mismatched*, and compression performance is degraded. This mismatch is complicated to model, and therefore for tractability it is assumed here that these scalar variances are known to the receiver. This could be achieved in practice by estimating the variance from a block of received symbols. It is known that when the assumed and true source statistics are close the performance loss due to mismatch is small [74].

The UQN rate allocation and quantization noise level can then be found using the bisection algorithm with

$$r_i = \log_2\Big(1 + \frac{\sigma_i^2}{\Delta}\Big). \tag{4.83}$$

This method can be easily generalised for scenarios where inter-cell users with unknown channels but known channel covariances introduce inter-cell interference, similar to how multi-cell signal processing is performed in [15].

Sum Capacity Under Imperfect CSI

By the reasoning in Section 2.4.2, the expected sum capacity of the transform coded MIMO uplink under imperfect CSI, is given by

$$\mathbb{E}_{\mathbf{E}}\big[\mathcal{C}_{\text{SUM}}\big] = \mathbb{E}_{\mathbf{E}}\Big[\log_2\det\Big(\mathbf{I}_K + \frac{\rho}{1+\Delta}\check{\mathbf{H}}^{\dagger}\check{\mathbf{H}}\Big)\Big] \tag{4.84}$$

$$= \mathbb{E}_{\mathbf{E}}\Big[\log_2\det\Big(\mathbf{I}_K + \frac{\rho}{1+\Delta}\hat{\mathbf{H}}^{\dagger}\Big(\mathbf{I}_M + \rho\sum_{k=1}^{K} p_k\mathbf{C}_k\Big)^{-1}\hat{\mathbf{H}}\Big)\Big]. \tag{4.85}$$

The capacity is now limited by three factors – quantization noise through Δ, channel estimation error through $\mathbf{C}_k$ & receiver noise through ρ.

The sum capacity has a monotonic relationship with SNR, ρ – increasing the user transmit power increases the sum capacity at any fronthaul rate and for any quality of CSI. However, since the channel estimation error and quantization noise components both scale with ρ, at high SNR

$$\lim_{\rho\to\infty}\ \mathbb{E}_{\mathbf{E}}\big[\mathcal{C}_{\text{SUM}}\big] = \mathbb{E}_{\mathbf{E}}\Big[\log_2\det\Big(\mathbf{I}_K + \frac{1}{1+\Delta}\hat{\mathbf{H}}^{\dagger}\Big(\sum_{k=1}^{K} p_k\mathbf{C}_k\Big)^{-1}\hat{\mathbf{H}}\Big)\Big], \tag{4.86}$$

the capacity is limited by both quantization noise and channel estimation error, and the fronthaul efficiency that is achievable therefore depends on the quality of CSI available to the system. Figure 4.18 shows this for i.i.d Rayleigh fading for transmit SNR $\rho = 10$ dB and a range of different uplink channel estimation pilot SNRs, ρ_{CSI} (cf. Section 2.4.2). With good quality CSI (large ρ_{CSI}), the behaviour is similar to the perfect CSI case, but when the CSI is of poor quality the fronthaul utilisation is poor. For example, 40 bpcu of fronthaul capacity gives a capacity of 39 bpcu for $\rho_{\text{CSI}} = 20$ dB, but only 22 bpcu for $\rho_{\text{CSI}} = 0$ dB. At high fronthaul capacities the quantization noise vanishes and it is the channel estimation error that limits performance.

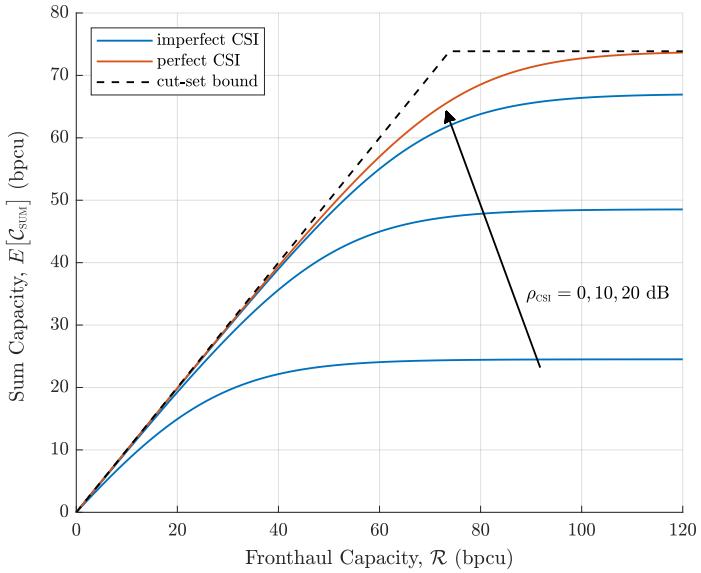

Figure 4.18: Sum capacity scaling with varying estimated CSI quality, i.i.d Rayleigh fading channel, $\rho = 10$ dB, $K = 8$, $M = 64$.

The impact of imperfect CSI is more significant at high SNR, ρ, as shown in Figure 4.19, since at low SNR the channel estimation error is small compared to receiver noise.

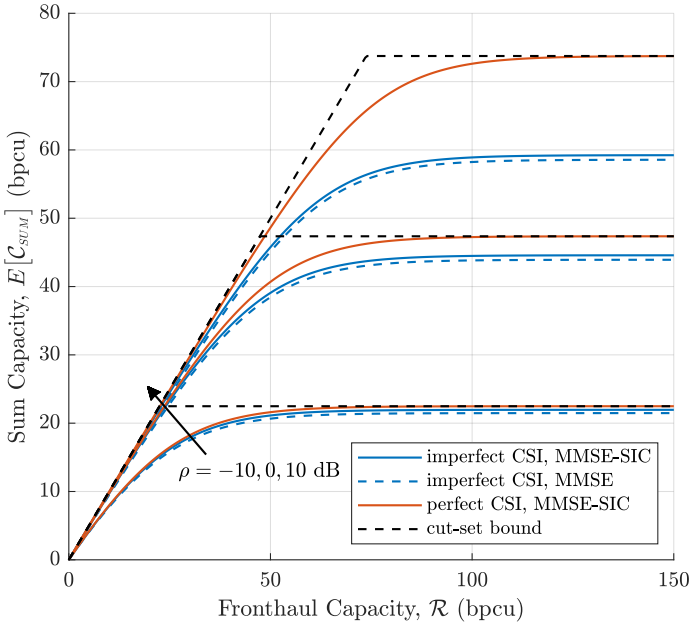

Figure 4.19: Sum capacity scaling with varying SNR, ρ, i.i.d Rayleigh fading channel, $\rho_{\text{CSI}} = 15$ dB, $K = 8$, $M = 64$.

4.4 Distributed MIMO Uplink Signal Compression

The application of transform coding to the distributed MIMO C-RAN topology shown in Figure 4.2 is now investigated. In this topology, there are now L receivers, each equipped with M antennas, that receive the uplink transmissions from all K users,

$$\mathbf{y}_l = \mathbf{H}_l \mathbf{x} + \boldsymbol{\eta}. \tag{4.87}$$

The receivers independently compress their received signals for transmission over individual fronthaul links with capacity $\mathcal{R}$ bpcu, using a linear transform and a set of scalar compressors with appropriately allocated compression rates,

$$\tilde{\mathbf{z}}_l = \mathbf{F}_l^{\dagger} \mathbf{y}_l + \boldsymbol{\delta}_l. \tag{4.88}$$

The uplink user symbols are then jointly detected using the set of L compressed signals.

Achieving efficient transform coding compression in a distributed MIMO setting requires more than simply replicating the single receiver scheme, because the received uplink signals at the L receivers are *dependent*. These dependencies must be accounted for through the use of global CSI if the transform coding is to achieve good fronthaul efficiency. Unfortunately, the problem of finding the L transforms and rate allocations that maximise sum capacity is non-convex, and – unlike the single receiver case – has no closed form solution.

This sections begins by showing the limitations of using transform coding with a UQN rate allocation in a distributed MIMO setting, before briefly outlining the successive convex approximation approach to jointly finding the optimal transforms and rate allocations for all receivers. A reduced complexity scheme that uses the KLT and globally calculated rate allocation is then

derived and analysed.

4.4.1 Transform Coding with a Uniform Quantization Noise Level

The transform coded signal at receive l is given by

$$\tilde{\mathbf{z}}_l = \mathbf{F}_l^\dagger \mathbf{H}_l \mathbf{x} + \mathbf{F}_l^\dagger \boldsymbol{\eta} + \boldsymbol{\delta}_l, \tag{4.89}$$

with diagonal quantization noise covariance, $\boldsymbol{\delta}_l \sim \mathcal{CN}\big(0, \boldsymbol{\Phi}_l\big)$. Placing no restriction on the number of antennas at each receiver (so that M may be less than or greater than K), the data signal component now lies in a $t = \min(M, K)$-dimensional subspace. It can now be assumed without loss of generality that $\tilde{\mathbf{z}}_l$ has t components, i.e. $\mathbf{F} \in \mathbb{C}^{M \times t}$ is semi-orthogonal.

Assuming the compressed signals are jointly used to detect the symbols under MMSE-SIC symbol detection, the sum capacity is given by

$$\begin{aligned} \mathcal{C}_{\text{SUM}} &= \mathcal{I}\big(\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_L; \mathbf{x}\big) && (4.90)\\ &= \log_2 \det \Big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \mathbf{H}_l^\dagger \mathbf{F}_l \big(\boldsymbol{\Phi}_l + \mathbf{I}_t\big)^{-1} \mathbf{F}_l^\dagger \mathbf{H}_l \Big) && (4.91)\\ &= \log_2 \det \Big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \mathbf{H}_l^\dagger \big(\boldsymbol{\Psi}_l + \mathbf{I}_M\big)^{-1} \mathbf{H}_l \Big), && (4.92) \end{aligned}$$

where $\boldsymbol{\Psi}_l$ is a non-diagonal quantization noise covariance matrix with a similar structure to (4.48).

In [248] it is shown that at high signal-to-quantization-plus-noise (SQNR), i.e when $\mathcal{R}$ and ρ are sufficiently large that

$$\rho \mathbf{H}_l \mathbf{H}_l^\dagger \gg \boldsymbol{\Psi}_l + \mathbf{I}_M, \tag{4.93}$$

then a uniform quantization noise level is approximately optimal. The sum capacity is then

$$\mathcal{C}_{\text{SUM}} = \log_2 \det \Big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \frac{\mathbf{H}_l^\dagger \mathbf{H}_l}{\Delta_l + 1} \Big), \tag{4.94}$$

where Δ_l at each receiver can be simply found using the bisection approach in Algorithm 2 to give

$$\sum_{i=1}^{t} \log_2 \Big(1 + \frac{\rho \lambda_{l,i} + 1}{\Delta_l} \Big) = \mathcal{R}, \tag{4.95}$$

where $\lambda_{l,i}$ are the t non-zero eigenvalues of $\mathbf{H}_l \mathbf{H}_l^\dagger$.

This scheme is attractive for its simplicity, but the high SQNR assumption is of limited interest since it describes the region where the impact of quantization noise is small; when the system operates with limited fronthaul capacity the high SQNR condition is not fulfilled.

A sum capacity upper bound can be established to assess the limits of performance under UQN compression in this fronthaul-limited region. This uses the lower bound on UQN

quantization noise

$$\sum_{i=1}^{t} \log_2\left(\frac{\rho\lambda_{l,i}}{\Delta_l}\right) \leq \mathcal{R} \implies \Delta_l \geq \rho\bar{\lambda}_l 2^{-\mathcal{R}/t}, \tag{4.96}$$

where $\bar{\lambda}_l$ is the geometric mean of the $\lambda_{l,i}$. Using this quantization noise lower bound, which is tight at high SNR, similarly to Section 4.3.1 an upper bound on capacity can be found,

$$\mathcal{C}_{\text{SUM}}^{\text{UB}} = \frac{RK}{t} + \log_2 \det\left(\sum_{l=1}^{L} \bar{\lambda}_l^{-1}\mathbf{H}_l^{\dagger}\mathbf{H}_l\right) + 2^{-\mathcal{R}/t}\log_2(e)\,\text{Tr}\left(\left(\sum_{l=1}^{L} \bar{\lambda}_l^{-1}\mathbf{H}_l^{\dagger}\mathbf{H}_l\right)^{-1}\right) \tag{4.97}$$

$$\approx \frac{RK}{t} + \epsilon, \tag{4.98}$$

where the approximation holds since the third term rapidly decays with $\mathcal{R}$. This effectively limits the fronthaul efficiency that can be achieved, as shown in Figure 4.20. In the i.i.d Rayleigh channel with a total fronthaul capacity of 80 bpcu, a sum capacity of 40 bpcu is achieved with $\rho = 10$ dB, and no more than 48 bpcu can be achieved at any SNR. For non-degraded MIMO

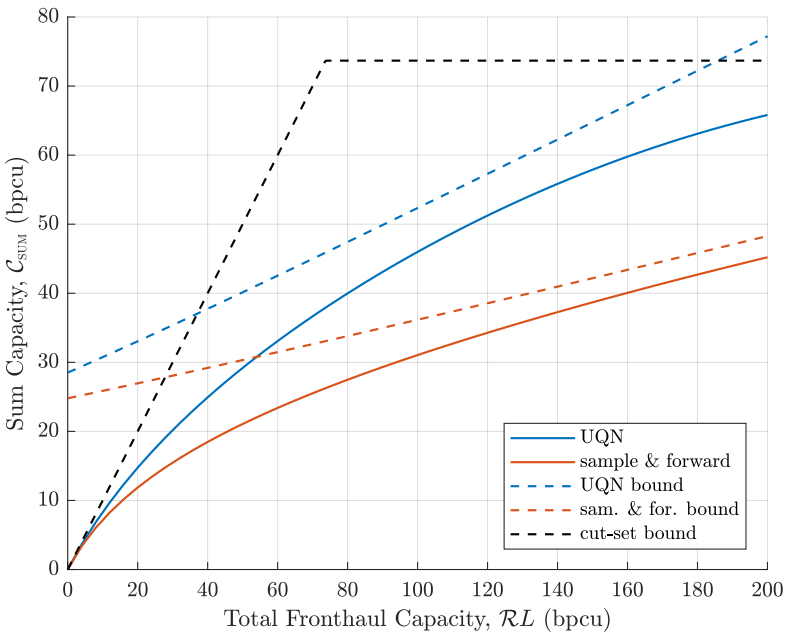

Figure 4.20: Distributed MIMO sum capacity under UQN compression, $\rho = 10$ dB, i.i.d Rayleigh fading channel, $K = 8$, $L = 4$ $M = 16$.

operation, it is required that $ML \geq K$, and hence necessarily

$$\frac{\mathcal{R}K}{t} \leq \mathcal{R}L \tag{4.99}$$

If an overall excess of antennas is deployed in the network in order to capture the benefits of channel hardening, $ML \gg K$, then $\mathcal{R}K/t \ll \mathcal{R}L$, and the sum capacity increases very slowly compared to the fronthaul capacity.

It should be noted that at low SNR the bound is loose and gives little information about the performance of UQN (sitting above the cut-set bound at some low SNRs). However, recall

that in 4.3.2 the local rate allocation was shown to result in poor fronthaul efficiency even in a single receiver setting at low SNR.

Thus, overall, whilst UQN transform coding compression is quasi-optimal in a setting with a *single* receiver and an excess of antennas, it performs poorly when these antennas are distributed between *multiple* receivers. Fundamentally, this is because transform coding with the UQN rate allocation does not account for dependencies between the signals received at different receivers. It can exploit *local* sparsity – reducing the number of signal components from M to K when $M \gg K$ – but does not exploit the *joint* sparsity that exists at a network level when $ML \gg K$. This can be seen in Figure 4.20, where UQN does not fully utilise the fronthaul, but does improve on sample & forward by accounting for local signal correlations.

4.4.2 Optimal Transform Coding for the Distributed MIMO Uplink

The sum capacity maximising transform coding scheme is found by optimising the quantization noise covariances, $\mathbf{\Psi}_l$, subject to the fronthaul rate constraints, $\mathcal{I}(\tilde{\mathbf{z}}_l; \mathbf{z}_l) \leq \mathcal{R}$,

$$\begin{aligned} \underset{\mathbf{\Psi}_l}{\text{maximise}} \quad & \log_2 \det \left(\mathbf{I}_K + \rho \sum_{l=1}^{L} \mathbf{H}_l^\dagger \left(\mathbf{\Psi}_l + \mathbf{I}_M \right)^{-1} \mathbf{H}_l \right) \\ \text{subject to} \quad & \log_2 \det \left(\rho \mathbf{H}_l \mathbf{H}_l^\dagger + \mathbf{I}_M + \mathbf{\Psi}_l \right) - \log_2 \det \left(\mathbf{\Psi}_l \right) \leq \mathcal{R} \quad \forall l. \end{aligned} \tag{4.100}$$

This is also the optimal point-to-point compression scheme – outperforming all compression schemes that do not make use of Wyner-Ziv-type distributed source coding. Unfortunately both its objective function and constraint are non-convex in the $\mathbf{\Psi}_l$, making finding a global maximum infeasible.

In [248], a method for finding a stationary point to the problem using successive convex approximation (SCA) is outlined. The full method is omitted here for space, but the basic idea is to convert the problem into a sequence of convex problems, by:

- replacing the objective function with a concave lower bound. This involves exploiting parallels between the sum capacity expression and MMSE detection.
- replacing the constraint with a convex upper bound.
- solving the resulting convex approximation of the problem using numerical methods, e.g the interior point method.
- updating the bounds such that at the current solution they are equal to the original functions, and re-solving, repeating until convergence.

This method finds a stationary point to the point-to-point compression sum capacity maximisation problem, and is referred to herein as SCA-P2P. Figure 4.21 shows the benefits of jointly optimising the quantization noise in a system with $K = 8$ users, $L = 4$ distributed receivers each equipped with $M = 8$ antennas, and SNR $\rho = 10$ dB. In the fronthaul-limited regime, a significant capacity improvement over UQN compression is experienced – for a fronthaul capacity of 60 bpcu, a sum capacity of 48 bpcu can be achieved, compared to 34 bpcu under UQN compression.

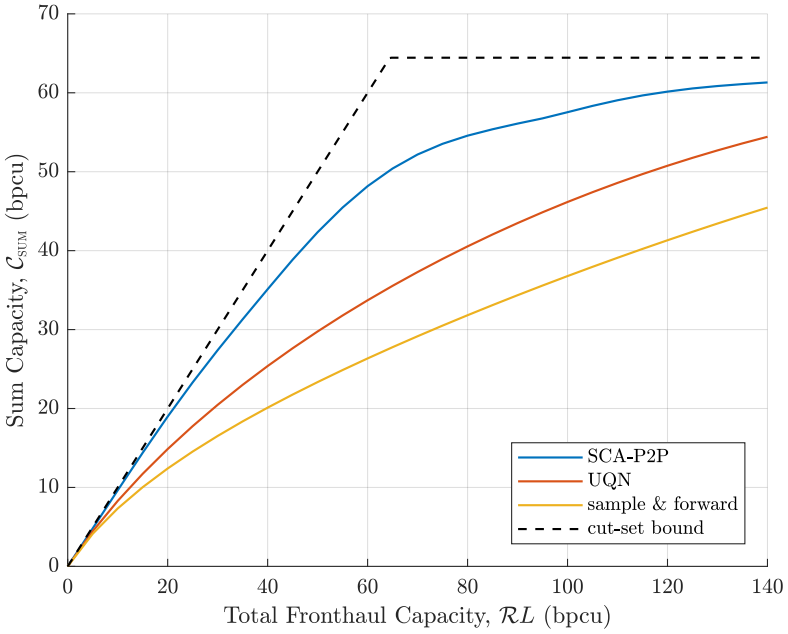

Figure 4.21: Distributed MIMO sum capacity with optimal point-to-point compression, $\rho = 10$ dB, i.i.d Rayleigh fading channel, $K = 8$, $L = 4$ $M = 8$.

However, the computational complexity of SCA-P2P is high due to the need to repeatedly solve convex problems using numerical solvers – prohibiting its use in large networks.

4.4.3 Transform Coding with Jointly Optimised Rate Allocation

As a more tractable alternative to jointly optimising the full quantization noise covariance matrix at all receivers, the optimisation can be a carried out over only the rate allocation, whilst using a fixed transform. The transforms can be separately calculated at each receiver using only local CSI (e.g the KLT), whilst the compression rates are allocated centrally at the CP using global CSI. This simplifies the optimisation problem from optimising L separate $t \times t$ complex covariance matrices under non-convex constraints to optimising Lt real valued scalars with simple linear constraints. Furthermore, from a practical perspective only t scalar values need to be transferred from the CP back to each receiver compared to $t(t+1)/2$ complex values each for the full covariance matrix, reducing signalling overheads.

A similar idea to this was previously investigated in [123] for maximising minimum user capacity under fixed-rate scalar quantizers, and shown to achieve good compression performance. However, the proposed solution involves numerically solving a large number of convex feasibility problems using the interior point method, leading to high computational complexity. With this in mind and in the interest of maximising fronthaul efficiency, here maximisation of the sum

capacity under either MMSE-SIC or MMSE detection is considered,

$$\underset{r_{l,i}}{\text{maximise}} \quad \mathcal{C}_{\text{SUM}} \quad \text{or} \quad \sum_{k=1}^{K} \mathcal{C}_k^{(\text{MMSE})} \tag{4.101}$$

$$\text{subject to} \quad \sum_{i=1}^{t} r_{l,i} \leq \mathcal{R}. \tag{4.102}$$

These problems have non-convex objectives, and hence cannot be directly solved to find the global maximum. However, inspired by the approach in [248], a successive convex approximation approach can be used to find a stationary point. It is shown here that this problem can be solved by iteratively updating the MMSE detection matrices and performing a rate allocation, where both problems have closed form solutions, and hence the use of computationally intensive numerical solvers is avoided.

Sum Capacity Maximisation

The basis of the successive convex approximation approach in [248] is to exploit the relationship between Gaussian capacity and MMSE detection,

$$\mathcal{C}_{\text{SUM}} = \log_2(\rho) + \log_2 \det\left(\mathbf{C}_e^{-1}\right) \tag{4.103}$$

where $\mathbf{C}_e$ is the error covariance matrix under MMSE detection,

$$\mathbf{C}_e = \rho\Big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \mathbf{H}_l^\dagger \mathbf{F}_l \big(\mathbf{I}_M + \mathbf{\Phi}_l\big)^{-1} \mathbf{F}_l^\dagger \mathbf{H}_l\Big)^{-1} \tag{4.104}$$

$$= \mathbb{E}\big[\big\|\mathbf{x} - \sum_{l=1}^{L} \mathbf{W}_l \tilde{\mathbf{z}}_l\big\|^2\big] \tag{4.105}$$

$$= \rho\Big(\mathbf{I}_K - \sum_{l=1}^{L} \mathbf{W}_l \mathbf{F}_l^\dagger \mathbf{H}_l\Big)\Big(\mathbf{I}_K - \sum_{l=1}^{L} \mathbf{W}_l \mathbf{F}_l^\dagger \mathbf{H}_l\Big)^\dagger + \sum_{l=1}^{L} \mathbf{W}_l \big(\mathbf{I}_M + \mathbf{\Phi}_l\big) \mathbf{W}_l^\dagger \tag{4.106}$$

where $\mathbf{W}_l$ is the MMSE detection matrix,

$$\mathbf{W}_l = \mathbf{C}_e \mathbf{H}_l^\dagger \mathbf{F}_l \big(\mathbf{I}_t + \mathbf{\Phi}_l\big)^{-1}, \tag{4.107}$$

and $\mathbf{F}_l \in \mathbb{C}^{M \times t}$ is a fixed (pre-selected) transform. In the numerical results provided here the local KLT as defined in Section 4.3.2 is used as the transform at each receiver, but the method is valid for any $\mathbf{F}_l \in \mathbb{C}^{M \times t}$ that is unitary/semi-orthogonal. Note that the formulation of the error matrix here involves the inversion of $K \times K$ matrix. This is equivalent (through the matrix inverse lemma) but preferable to the conventional MMSE error matrix formulation that would involve the inversion of a $Lt \times Lt$ matrix, since $Lt \geq K$.

Using this relationship, the successive convex approximation rate allocation procedure (SCA-RA) finds a stationary point to the optimal rate allocation problem by:

1. **Replacing the maximisation with an equivalent non-convex minimisation.**

$$\underset{r_{l,i}}{\text{maximise}} \quad \mathcal{C}_{\text{SUM}} \implies \underset{r_{l,i}}{\text{minimise}} \quad \log_2 \det\left(\mathbf{C}_e\right). \tag{4.108}$$

2. **Approximating the non-convex objective function with a convex upper bound,** using the identity

$$\log_2 \det\left(\mathbf{C}_e\right) \leq \log_2 \det\left(\mathbf{A}\right) + \text{Tr}\left(\mathbf{A}\mathbf{C}_e\right) - K \tag{4.109}$$

which achieves equality when

$$\mathbf{A} = \mathbf{C}_e^{-1}. \tag{4.110}$$

3. **Finding the rate allocation that minimises this upper bound,** for a fixed MMSE detection matrix. Substituting in (4.106), and removing the components that do not depend on $r_{l,i}$, the optimisation separates into L rate allocation optimisations,

$$\begin{aligned} &\underset{r_{l,i}}{\text{minimise}} \quad \text{Tr}\left(\mathbf{W}_l^\dagger \mathbf{A}\mathbf{W}_l \mathbf{\Phi}_l\right) \\ &\text{subject to} \quad \sum_{i=1}^{t} r_{l,i} \leq \mathcal{R}. \end{aligned} \tag{4.111}$$

Since $\mathbf{\Phi}_l$ is diagonal,

$$\text{Tr}\left(\mathbf{W}_l^\dagger \mathbf{A}\mathbf{W}_l \mathbf{\Phi}_l\right) = \sum_{i=1}^{t} \frac{q_{l,i}}{2^{r_{l,i}} - 1}, \tag{4.112}$$

where

$$q_{l,i} = \left(\rho \mathbf{f}_{l,i}^\dagger \mathbf{H}_l \mathbf{H}_l^\dagger \mathbf{f}_{l,i} + 1\right) \times \left[\mathbf{W}_l^\dagger \mathbf{A}\mathbf{W}_l\right]_{i,i}. \tag{4.113}$$

Fixing the $\mathbf{W}_l$, and formulating the Lagrangian, this has a simple closed form solution (see Appendix 2.2),

$$r_{l,i} = \left[\log_2\left(\gamma_l + q_{l,i} + \sqrt{q_{l,i}}\sqrt{q_{l,i} + 2\gamma_l}\right) - \log_2(\gamma_l)\right]^+ \tag{4.114}$$

where $\gamma_l \in \mathbb{R}^+$ is found such that the fronthaul constraint is met, using a similar bisection approach to Algorithm 2.

4. **Updating the convex approximation,** so that it has equality with the non-convex objective at the current rate allocation. Using the updated rate allocation from step 3, the updated quantization noise covariance matrices, $\mathbf{\Phi}_l$, can be calculated, and an updated convex problem formed by updating $\mathbf{C}_e$ as in (4.106), $\mathbf{A}$ as in (4.110), and $\mathbf{W}_l$ as in (4.107). By substitution these steps can be absorbed into the simple updates,

$$\phi_{l,i}^\star = \left[\mathbf{\Phi}_l^\star\right]_{i,i} = \frac{\rho \mathbf{f}_{l,i}^\dagger \mathbf{H}_l \mathbf{H}_l^\dagger \mathbf{f}_{l,i} + 1}{2^{r_{l,i}} - 1}, \tag{4.115}$$

and

$$q^{\star}_{l,i} = \left(\rho \mathbf{f}^{\dagger}_{l,i} \mathbf{H}_l \mathbf{H}^{\dagger}_l \mathbf{f}_{l,i} + 1\right) \times \frac{\rho}{(1+\phi^{\star}_{l,i})^2} \mathbf{f}^{\dagger}_l \mathbf{H}_l \Big(\mathbf{I}_K + \rho \sum_{j=1}^{L} \mathbf{H}^{\dagger}_j \mathbf{F}_j (\mathbf{I}_M + \mathbf{\Phi}^{\star}_l) \mathbf{F}^{\dagger}_j \mathbf{H}_j \Big)^{-1} \mathbf{H}^{\dagger}_l \mathbf{f}_l. \tag{4.116}$$

5. **Repeating steps 3 & 4 until convergence.** The rates allocated in step 3 meet the constraints of the new updated convex problem calculated in step 4, but do not generally minimise it. Solving for a new rate allocation with the updated $q_{l,i}$ therefore monotonically decreases the objective function, and when applied iteratively this procedure converges monotonically towards a sum capacity maximising (stationary point) rate allocation.

The full algorithm is shown in Algorithm 3.

Algorithm 3 SCA-RA Sum Capacity Maximising Rate Allocation

inputs: $\mathbf{H}_l, \mathbf{F}_l, \rho, \mathcal{R}$

initialise $r_{l,i}$

initialise $\phi_{l,i} \leftarrow \dfrac{\rho \mathbf{f}^{\dagger}_{l,i} \mathbf{H}_l \mathbf{H}^{\dagger}_l \mathbf{f}_{l,i} + 1}{2^{r_{l,i}} - 1}$

initialise $\mathbf{C}_e \leftarrow \rho \Big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \sum_{i=1}^{t} \frac{\mathbf{H}^{\dagger}_l \mathbf{f}_{l,i} \mathbf{f}^{\dagger}_{l,i} \mathbf{H}_l}{1+\phi_{l,i}} \Big)^{-1}$

while $\Delta \mathcal{C}_{\text{SUM}} \geq \epsilon$ **do**

  for $l = 1 : L$ **do**

    $q_{l,i} \leftarrow \left(\rho \mathbf{f}^{\dagger}_{l,i} \mathbf{H}_l \mathbf{H}^{\dagger}_l \mathbf{f}_{l,i} + 1\right) \dfrac{\mathbf{f}^{\dagger}_l \mathbf{H}_l \mathbf{C}_e \mathbf{H}^{\dagger}_l \mathbf{f}_l}{(1+\phi_{l,i})^2}$

    $r_{l,i} \leftarrow \Big[\log_2 \left(\gamma_l + q_{l,i} + \sqrt{q_{l,i}} \sqrt{q_{l,i} + 2\gamma_l} \right) - \log_2(\gamma_l) \Big]^{+} \qquad \gamma_l : \sum_{i=1}^{t} r_{l,i} = \mathcal{R}$

  end for

  $\phi_{l,i} \leftarrow \dfrac{\rho \mathbf{f}^{\dagger}_{l,i} \mathbf{H}_l \mathbf{H}^{\dagger}_l \mathbf{f}_{l,i} + 1}{2^{r_{l,i}} - 1}$

  $\mathbf{C}_e \leftarrow \rho \Big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \sum_{i=1}^{t} \dfrac{\mathbf{H}^{\dagger}_l \mathbf{f}_{l,i} \mathbf{f}^{\dagger}_{l,i} \mathbf{H}_l}{1+\phi_{l,i}} \Big)^{-1}$

  $\mathcal{C}_{\text{SUM}} \leftarrow \log_2(\rho) - \log_2 \det \left(\mathbf{C}_e \right)$

end while

outputs: $r_{l,i}, \phi_{l,i}$

The computational complexity of the optimal rate allocation method is dominated at each iteration by the matrix inversion, with overall complexity $\mathcal{O}\big(K^3 N_{\text{it}}\big)$, where N_{it} iterations are required for convergence. Since the SCA-RA method does not require the use of any numerical solvers, its execution time was found to be many orders of magnitude shorter than SCA-P2P. A further discussion on computational complexity is provided in Section 4.4.6.

Fixing the transform incurs a performance penalty over SCA-P2P, since the algorithm has the freedom to optimise the number of bits allocated to each dimension of the signal subspace, but not the freedom to optimise the signal compression basis. However, numerical results indicate that this performance loss is small, and most of the benefits over UQN compression are

maintained. For example, Figure 4.22 compares the performance for $K = 8, L = 8, M = 4$ at low SNR, $\rho = 0$ dB. This indicates that the main benefit of jointly optimising the quantization noise covariance matrices comes from jointly determining how much quantization noise to introduce onto different signal dimensions, rather than from finding the optimal signal basis.

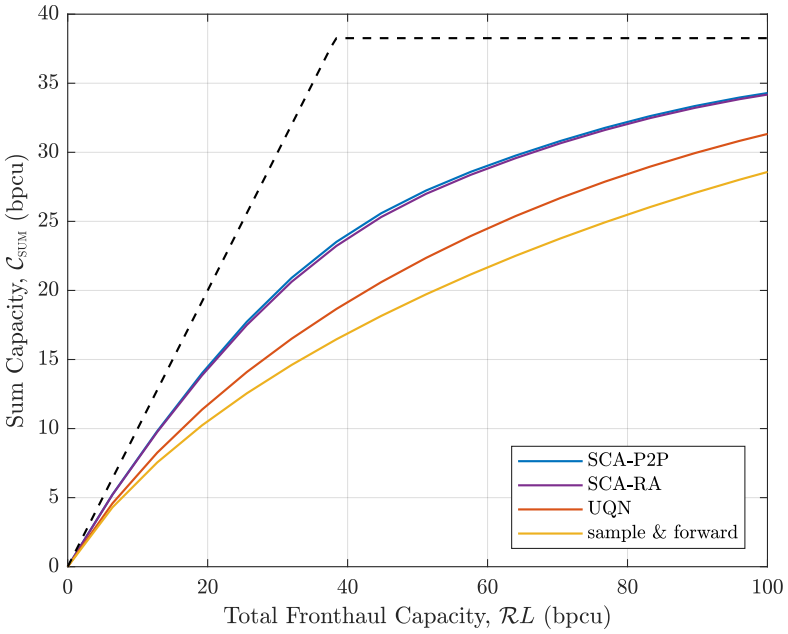

Figure 4.22: Distributed MIMO sum capacity with KLT and optimal SCA rate allocation, i.i.d Rayleigh fading channel, $K = 8$, $L = 8$ $M = 4$, $\rho = 0$ dB.

Capacity Maximisation under Linear Detection

Under MMSE detection, the capacity of user K is

$$\mathcal{C}_k = \log_2(\rho) - \log_2(e_k) \tag{4.117}$$

where $e_k = \left[\mathbf{C}_e\right]_{k,k}$ is the MMSE error covariance for symbol k. The sum user capacity maximisation can be written

$$\underset{r_{l,i}}{\text{maximise}} \quad \sum_{k=1}^{K} \mathcal{C}_k \implies \underset{r_{l,i}}{\text{minimise}} \quad \sum_{k=1}^{K} \log_2(e_k), \tag{4.118}$$

and the convex upper bound,

$$\log_2(e_k) \leq \log_2(a_k) + a_k e_k - 1, \tag{4.119}$$

which has equality when $a_k = 1/e_k$, can be used to approximate the non-convex problem as

$$\underset{r_{l,i}}{\text{minimise}} \quad \text{Tr}\left(\mathbf{W}_l^{\dagger} \tilde{\mathbf{A}} \mathbf{W}_l \mathbf{\Phi}_l\right), \tag{4.120}$$

where $\tilde{\mathbf{A}} = \mathrm{diag}(a_k)$. The rate allocation that maximises sum capacity under linear detection is therefore found using an identical procedure to that under optimal detection, but replacing $q_{l,i}$ with

$$\tilde{q}_{l,i} = \left(\rho \mathbf{f}_{l,i}^{\dagger} \mathbf{H}_l \mathbf{H}_l^{\dagger} \mathbf{f}_{l,i} + 1\right) \times \left[\mathbf{W}_l^{\dagger} \tilde{\mathbf{A}} \mathbf{W}_l\right]_{i,i}. \tag{4.121}$$

As seen in Figure 4.23, linear detection results in some sum capacity loss compared to optimal detection, but the joint rate allocation ensures improved capacity scaling compared to local UQN compression.

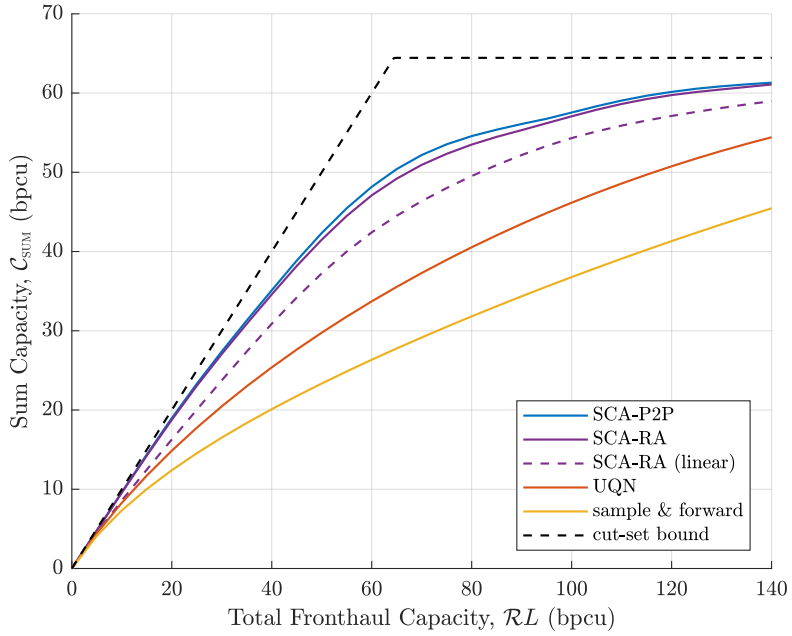

Figure 4.23: Distributed MIMO sum capacity with KLT and optimal SCA rate allocation, i.i.d Rayleigh fading channel, $\rho = 10$ dB, $K = 8$, $L = 4$ $M = 8$.

Sparse Rate Allocation

An interesting feature of the SCA rate allocation when operating in the fronthaul-limited region is that when there are an overall excess of receive antennas, $ML \gg K$, the SCA rate allocation tends to be sparse – many of the $r_{l,i}$ are zero or near-zero, and only a subset of the available signal components at each receiver are compressed. This effect is most pronounced at high SNR, and as shown in Figure 4.24.

This was previously observed in [123], but not investigated in detail. Here, from Algorithm 3 the specific mechanism that produces the sparse rate allocation can be identified:

- When the factor $q_{l,i}$ is small, $z_{l,i}$ will be allocated a small number of compression bits as in (4.114).
- With a small number of compression bits, $r_{l,i}$, allocated, $\tilde{z}_{l,i}$ will contain a larger amount of quantization noise, (4.115).

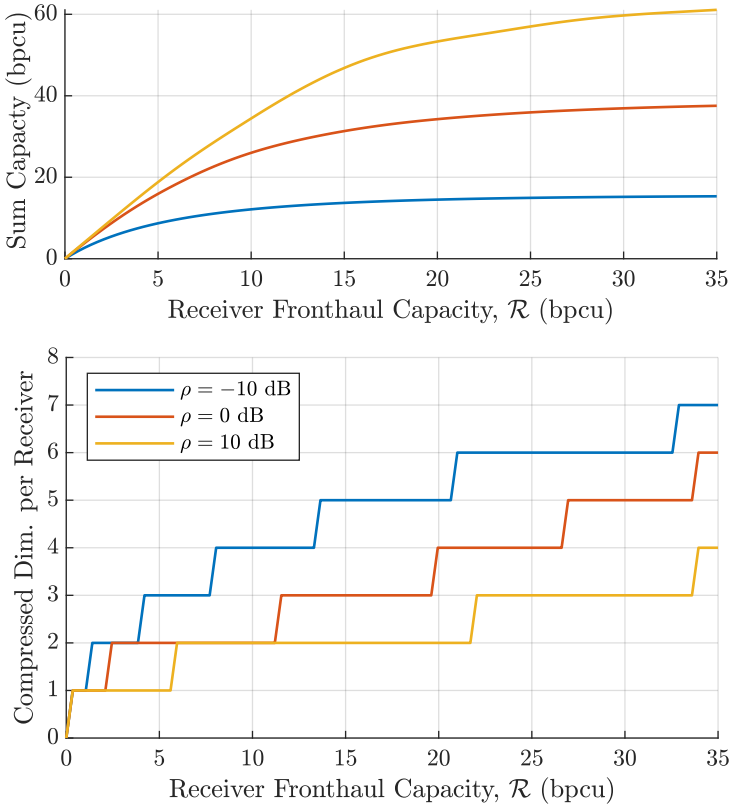

Figure 4.24: Average (mode) number of signal components allocated $r_{l,i} \geq 0.1$ bpcu at different fronthaul capacities, i.i.d Rayleigh fading channel, $K = 8$, $M = 8$, $L = 4$.

- When the MMSE combining matrix is updated, $\tilde{z}_{l,i}$ will then have a smaller weighting in the MMSE estimate due to its high quantization noise.
- This low MMSE weighting will reduce $q_{l,i}$ at the next stage, (4.116).
- Applied iteratively, this will lead to some of the signal components being allocated zero (or near-zero) rates. The final rate allocation sparsity pattern depends on the available fronthaul capacity, SNR & channel realisation.

The SCA-RA algorithm can be interpreted as implictly performing dimension reduction on the transformed signal, in response to the joint sparsity of the received signals when $ML \gg K$. The explicit use of dimension reduction for signal compression is the subject of Chapter 5, where this idea is explored in more detail.

4.4.4 Distributed MIMO with Limited Fronthaul

Having showed that transform coding with locally calculated KLT transforms and a jointly optimised rate allocation significantly outperforms transform coding with a local rate allocation, some further insights into the scheme's performance in distributed MIMO networks are now provided using numerical examples.

High SNR Behaviour

As the SNR is increased the overall fronthaul efficiency increases, as shown in Figure 4.25. As seen above, in the fronthaul-limited region the SCA-RA scheme will tend to select a reduced number of signal components and allocate the compression bits to these. At high SNR, when the selected components contain relatively less noise, the performance is limited mainly by quantization noise, improving fronthaul efficiency. However, unlike the single receiver case, the SCA-RA scheme will generally not fully achieve the cut-set bound at high SNR.

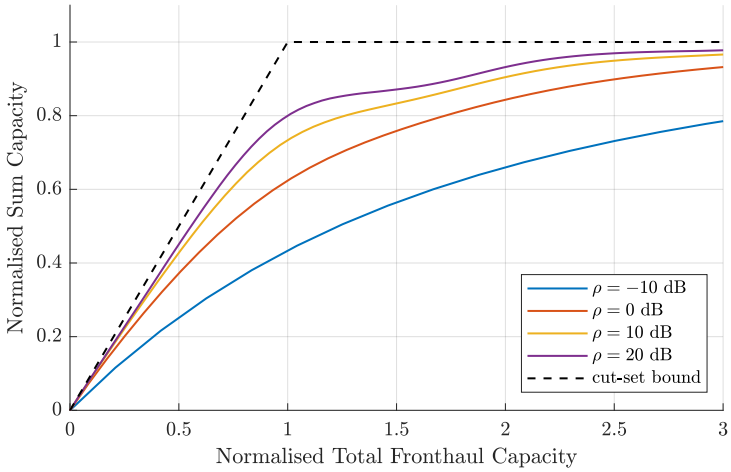

Figure 4.25: Fronthaul utilisation, $K = 8$, $M = 8$, $L = 4$.

In a dense distributed MIMO deployment, as simulated here, operation in the high SNR regime (where each user is received by at least one receiver with high SNR) is a realistic assumption due to the reduced pathloss/macro-diversity provided by distributing receivers [234].

Performance as Network Density Increases

Figure 4.26 shows the mean user throughputs achieved in a 100 MHz channel, assuming overheads of 20%, i.e.

$$\text{mean user spectral efficiency} = 0.8 \times \text{mean user capacity}. \tag{4.122}$$

Fixing the number of antennas per receiver, M, overall ratio of antennas to users, $ML/K = 4$, and applying power control such that the total average received power for each user is kept constant (i.e. halving the transmit power of each user when the number of receivers is doubled[10]), the joint rate allocation ensures that good performance is maintained as the density of receivers and users is increased, by accounting for dependencies in the received signals.

The overall throughput benefits over local UQN compression and sample & forward compression are significant. For example, with 8 receivers each equipped with 8 antennas and separate 2 Gbps fronthaul connections, the SCA-RA scheme provides an additional 100 Mbps of throughput for each of 16 users relative to UQN compression.

[10]If the user transmit power is instead kept constant as the network density increases, the effective SNR of the transformed signals is increased, and user capacities increase.

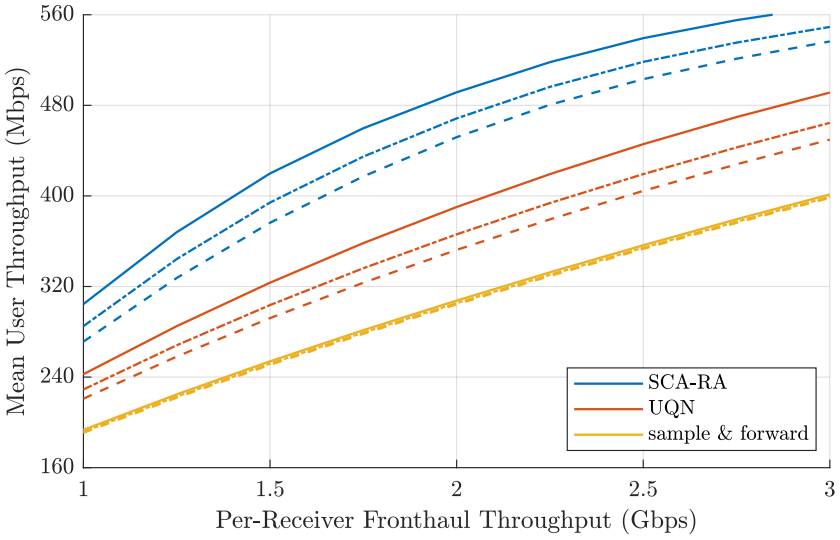

Figure 4.26: Mean user throughputs in 100 MHz channel for different user & RRH densities, i.i.d Rayleigh fading channel, $M = 8$. Solid line: $K = 8, L = 4, \rho = 10$ dB, dot-dash line: $K = 16, L = 8$, dashed line: $K = 32, L = 16$.

Benefit of Adding Antennas at the Receivers

Since boosting the SNR of the transformed signal improves fronthaul efficiency, increasing the number of antennas at each receiver allows either the user transmit power to be reduced or the user capacities & fronthaul efficiency to be increased (by the same reasoning as in Section 4.3.4), as shown in Figure 4.27. However, as the SNR increases the capacity benefit of increasing the number of antennas diminish, the user capacities being limited by the fronthaul capacity, as seen in the top subfigure.

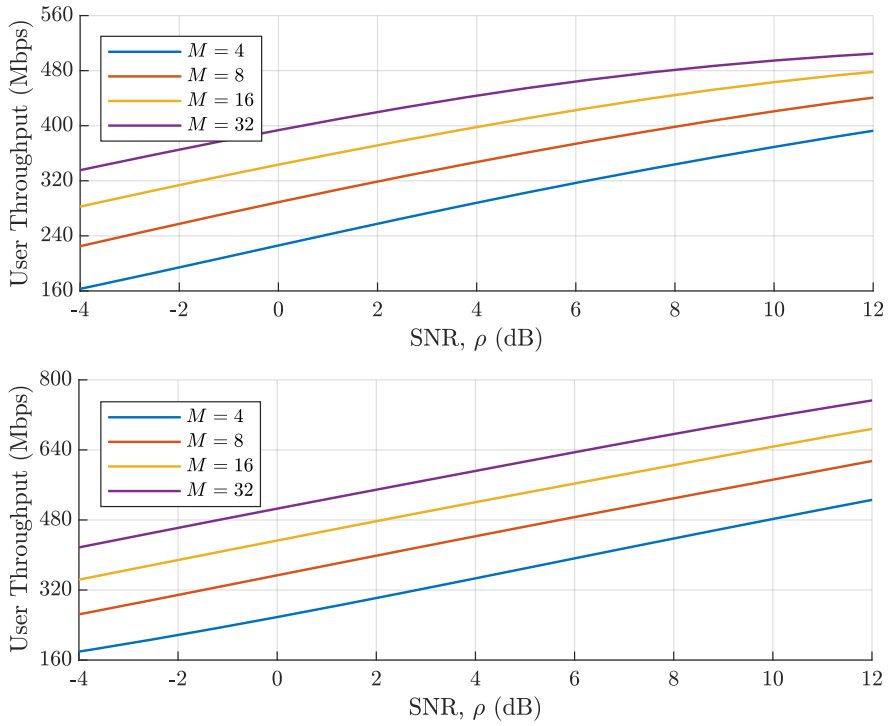

Figure 4.27: Distributed MIMO mean user throughputs in 100 MHz channel with varying numbers of receiver antennas, i.i.d Rayleigh fading channel, $K = 8, L = 4$. Top: per-receiver fronthaul throughput 1.5 Gbps, bottom: per-receiver fronthaul throughput 3 Gbps.

4.4.5 Transform Coding with Imperfect CSI

Applying the transform to the received signal under imperfect CSI as in Section 4.4.6, the expected sum capacity under imperfect CSI is

$$\mathbb{E}_{\mathbf{E}}\big[\mathcal{C}_{\text{SUM}}\big] = \mathbb{E}_{\mathbf{E}}\Big[\log_2 \det\Big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \check{\mathbf{H}}_l^{\dagger}\mathbf{F}_l\big(\mathbf{I}_t + \mathbf{\Phi}_l\big)^{-1}\mathbf{F}_l^{\dagger}\check{\mathbf{H}}_l\Big)\Big], \tag{4.123}$$

where, under the same assumptions as in Section 4.3.5,

$$\big[\mathbf{\Phi}_l\big]_{i,i} = \frac{\sigma_{l,i}^2}{2^{r_{l,i}} - 1} \tag{4.124}$$

with

$$\sigma_{l,i}^2 = \mathbb{E}\big[\big|\mathbf{f}_{l,i}^{\dagger}\mathbf{y}_l\big|^2\big] = \rho\mathbf{f}_{l,i}^{\dagger}\mathbf{\Omega}_l^{-1/2}\mathbf{H}_l\mathbf{H}_l^{\dagger}\mathbf{\Omega}_l^{-1/2}\mathbf{f}_{l,i} + \mathbf{f}_{l,i}^{\dagger}\mathbf{\Omega}_l^{-1}\mathbf{f}_{l,i}. \tag{4.125}$$

The MMSE detection matrix is

$$\mathbf{C}_e = \mathbb{E}_{\mathbf{E}}\big[\big\|\mathbf{x} - \sum_{l=1}^{L}\mathbf{W}_l\tilde{\mathbf{z}}_l\big\|^2\big] \tag{4.126}$$

$$= \rho\Big(\mathbf{I}_K - \sum_{l=1}^{L}\mathbf{W}_l\mathbf{F}_l^{\dagger}\check{\mathbf{H}}_l\Big)\Big(\mathbf{I}_K - \sum_{l=1}^{L}\mathbf{W}_l\mathbf{F}_l^{\dagger}\check{\mathbf{H}}_l\Big)^{\dagger} + \sum_{l=1}^{L}\mathbf{W}_l\big(\mathbf{I}_M + \mathbf{\Phi}_l\big)\mathbf{W}_l^{\dagger}, \tag{4.127}$$

and the rate allocation method can be immediately adapted to the case of imperfect CSI using

$$q_{l,i} = \sigma^2_{l,i} \times \left[\mathbf{W}_l^{\dagger}\mathbf{A}\mathbf{W}_l\right]_{i,i}. \tag{4.128}$$

Figure 4.28 shows the performance of the SCA-RA scheme with differing levels of CSI quality. When the channel estimates are good, the scheme performs close to the perfect case CSI, as expected, but when CSI is poor the fronthaul efficiency is reduced. When the CSI is poor the benefits of using the joint rate allocation instead of the local UQN compression also reduce.

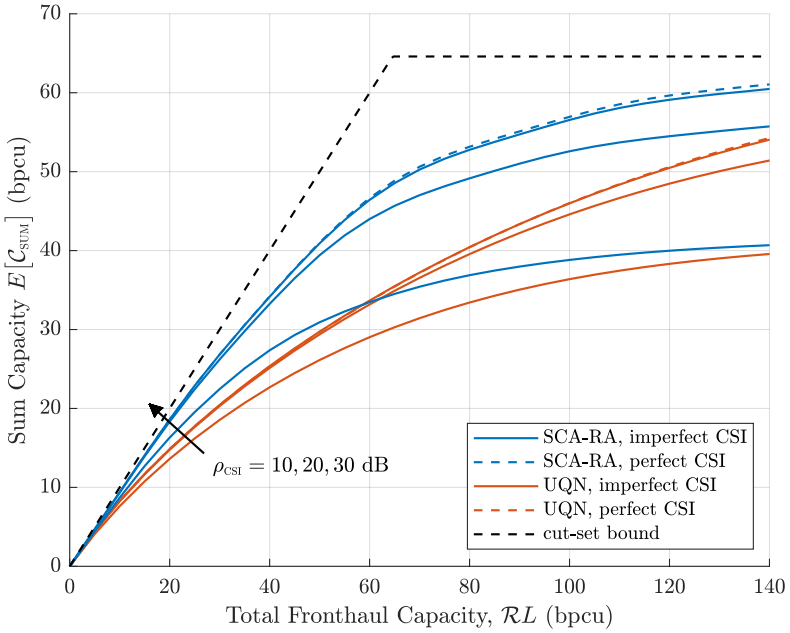

Figure 4.28: SCA-RA with imperfect CSI, i.i.d Rayleigh fading channel, $K = 8$, $L = 4$ $M = 8$, $\rho = 10$ dB.

4.4.6 Practical Aspects

This section considers the computational complexity and fronthaul signalling overheads associated with the scheme.

Computational Complexity

The SCA-RA algorithm uses CSI from all receivers, and therefore must be performed centrally at the CP. When the number of users is large, the computational complexity of Algorithm 3 is dominated by the matrix inversion required at each iteration, and scales as $\mathcal{O}\big(K^3 N_{\text{it}}\big)$. In general, the number of iterations, N_{it}, required for full convergence was found through testing to vary depending on the parameters $\mathcal{R}$, K, L, M and ρ. However, since the algorithm converges monotonically, a fixed number of iterations can be used to maximise performance under a computational complexity constraint. Numerical results indicate that, initialising SCA-RA using the UQN rate allocation, $N_{\text{it}} \sim 6$ is generally sufficient to capture most of the benefits of jointly optimising the rate allocation, as shown in Figure 4.29.

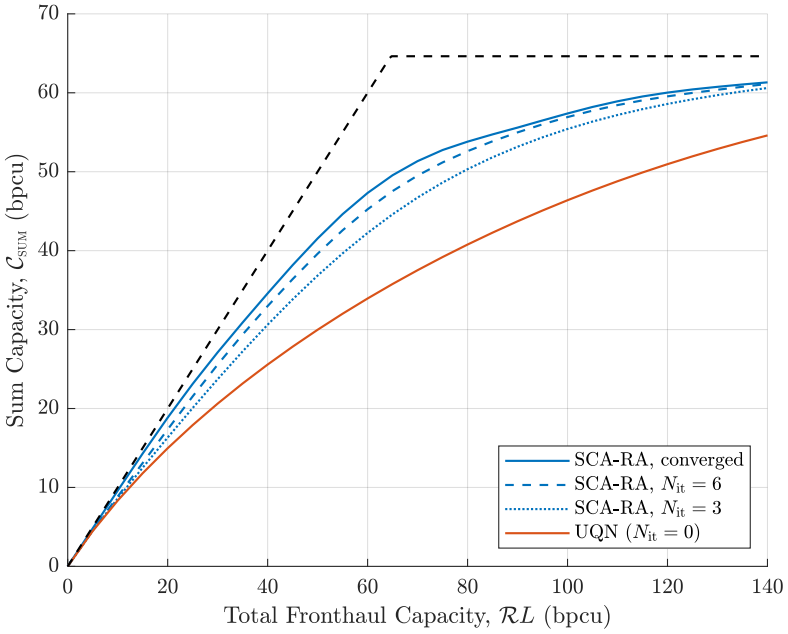

Figure 4.29: SCA-RA with limited number of iterations, $\rho = 10$ dB, i.i.d Rayleigh fading channel, $K = 8$, $L = 4$ $M = 8$.

This rate allocation must be performed once for each coherence block, and therefore the overall computational loads depends on the coherence times and bandwidths of the channels. The complexity of performing symbol detection under linear detection is $\mathcal{O}(KLt)$.

The calculation of each KLT requires the singular value decomposition of $\mathbf{H}_l\mathbf{H}_l^\dagger$ to be computed, which has complexity $\mathcal{O}(MK^2)$. This can be performed locally at the receiver, to reduce overheads, or at the CP to minimise computation at the distributed radio heads – see below.

Fronthaul Signalling Overheads

Since the rate allocation is performed at the CP but compression performed at the distributed receivers, there are various fronthaul signalling overheads associated with the SCA-RA scheme. Assuming the uplink channel matrices, $\mathbf{H}_l$, are initially obtained at the distributed receivers through channel estimation then the CP and receivers have the following requirements:

- The CP requires knowledge of the L composite transformed channels $\mathbf{F}_l^\dagger\mathbf{H}_l$ in order to perform rate allocation and then symbol detection. If the composite channel is known then the original channel matrix does not need to be known to the CP.
- The receivers require knowledge of their local transform, $\mathbf{F}_l$, and rate allocation, $r_{l,i}$ in order to perform compression.

As mentioned above, there are two options:

1. The L KLTs can be calculated locally at their respective receivers, and the composite channel, $\mathbf{F}_l^\dagger\mathbf{H}_l$, transferred to the CP. The composite channel has tK values whilst the channel matrix has MK values, and hence for $M > K = t$ this reduces the amount of

CSI transferred from each receiver to the CP. Only the t scalar values, $r_{l,i}$, need to be fed back to each receiver.

2. The L KLTs can be calculated at the CP, and then transferred back over fronthaul to the receivers. The CP then requires the original channel matrix, MK values, whilst the Mt values for the $\mathbf{F}_l$ need to be transferred back to each receiver, along with the t rate allocation values.

The first option results in lower overheads in both forward and reverse directions, but requires the receivers to have more computational power.

All of these overheads occur once per coherence block, and therefore the proportion of fronthaul capacity that must be reserved depends on the coherence times and bandwidths (reducing proportionally as the coherence block grows). In practice there may be opportunities for reducing overheads – here it is assumed that each user has a channel to each receiver, but for users far from a receiver the channels are very weak, and hence the effective channel matrix size and channel rank t is reduced.

Further investigation is required to study methods for reducing overheads and the implications for the achievable capacity of the system, including the channel coherence block size in real mobile channels, and methods for compressing the CSI and their impact on performance.

Practical Scalar Compression

This analysis has considered only the information-theoretic Gaussian scalar compression model. Whilst this is a commonly used tool for compression analysis in communication systems, it represents an approximation of real-life systems, since:

- the symbols used in practical communication systems come from non-Gaussian alphabets, and therefore whilst signals being compressed share the same second-order statistics, they are not truly Gaussian distributed. Designing and analysing optimal compression schemes for the true signal distributions is intractable when each signal consists of linear combinations of symbols coming from, for example, QAM alphabets. Thus in practice the Gaussian approximation (which by the central limited theorem is a reasonable approximation when the signal being compressed is the sum of many non-Gaussian symbols) can be used and some performance degradation due to quantizer mismatch accepted. Similarly, the resulting capacity expressions – as with all MIMO capacity analysis – are only approximations of the achievable performance when QAM symbols are used.

- the Gaussian scalar compression scheme upper bounds the performance of real life compression schemes, as discussed in Section 4.2.2. Thus real compression schemes will suffer from either a user data throughput penalty at a given fronthaul throughput, or a fronthaul throughput penalty at a given user data throughput. However, using, for example, entropy coded scalar quantization the rate penalty is small for each compressor, and thus under a sparse rate allocation the overall penalty is expected to be small. The entropy coding stage has complexity $\mathcal{O}(1)$ per scalar [180], but for minimal complexity fixed-rate scalar quantization could instead be used – at the expense of a larger performance penalty.

Since only approximate expressions are possible without making the above assumptions, future work should use bit-level simulations to benchmark the actual performance that can be achieved under the proposed scheme.

4.5 Conclusion

Limited capacity fronthaul connections based on wireless point-to-point or ethernet links will play an important role in enabling dense & flexible C-RAN deployments in future wireless networks. However, their use depends on the availability of bespoke data compression techniques for efficiently reducing the amount of data that must be transferred over fronthaul. This chapter has investigated the use of transform coding for compressing the multi-antenna received signals on the MIMO C-RAN uplink, giving particular attention to cases where the MIMO capacity is fundamentally limited by the available fronthaul capacity.

Massive MIMO

The first half of this chapter focuses on uplink signal compression for a single massive MIMO receiver. It is shown that due to the large number of antennas, individually compressing the received signal at each antenna gives very poor performance when the fronthaul capacity is limited – each signal must be compressed at very low resolution, resulting in high levels of quantization noise that limit the MIMO capacity. Transform coding, however, is shown to form a natural partner to massive MIMO – using a linear decorrelating transform to exploit the inherent sparsity in the received signal, reducing the number of signal components that need to be compressed.

Using the Karhunen-Loeve transform (KLT) in conjunction with a set of optimal Gaussian scalar compression codebooks, it is shown that transform coding uses the available fronthaul capacity very efficiently, especially at higher SNRs; at lower fronthaul capacities, where performance is limited by quantization noise, increasing the fronthaul capacity increases the achievable MIMO sum capacity by almost the same amount. For example, numerical results presented here show that if a receiver with 64 antennas, 8 users and SNR of 0 dB has 32 bits with which to compress its received signal (per subcarrier), a mean sum capacity of over 28 bits per channel use can be achieved.

This chapter shows that even at lower fronthaul capacities, where the MIMO capacity is unavoidably limited by quantization noise, many of the benefits associated with deploying a large number of antennas are present under transform coding – the effects of fast fading disappear, linear processing becomes optimal and the array gain enables transmit power to be reduced. The scheme is adapted for the case of imperfect CSI, with numerical results showing the importance of having good quality channel estimates.

As well as serving as a useful introduction into the use of transform coding in MIMO C-RAN networks, these findings suggest that transform coding can be an effective strategy for compressing the large quantity of sampled data produced on the massive MIMO uplink for transfer over fronthaul – capturing the channel hardening properties and energy efficiency gains of massive MIMO even when fronthaul capacity is limited. The complexity of performing

transform coding compression is low – requiring only the application of a linear transform, and then entropy-coded quantization of the transformed variables. The most computationally intensive task is calculating the KLT, which has complexity $\mathcal{O}(MK^2)$ and is only required once per coherence block.

However, with a single remote MIMO receiver, compressing the received signal and forwarding it to the CP for detection is not the optimal strategy – the symbol detection can instead be performed at the receiver, and the decoded user streams transferred over fronthaul (in which case the fronthaul can be utilised perfectly when the sum of the user capacities match the fronthaul capacity). The results in this chapter suggest that the transform coding 'compress & forward' strategy could be a viable alternative to 'detect & forward' for massive MIMO, and may be attractive to implement depending on the desired split of functionality between the remote unit and the CP. This is an important area for further investigation, with energy efficiency being a key consideration – the signals under detect & forward are not degraded by quantization noise, meaning for a given capacity lower user transmit power is required compared to compress & forward – this must be traded off against the computational resources required at the remote receiver.

Distributed MIMO

The second half of this chapter considers uplink signal compression for distributed MIMO. Here the multi-antenna signal at each receiver is compressed using transform coding before being transferred to the CP over its own fronthaul link. Since joint symbol detection must be performed at the CP using the full set of received signals, the compress & forward strategy is attractive in this context, with transform coding representing a low complexity solution.

First, this chapter shows that for efficient fronthaul utilisation the transform coding schemes must be jointly optimised across all receivers to account for correlations/dependencies between the received signals – an upper bound is established for the sum capacity achieved when there is no co-ordination between receivers. Whilst a procedure for jointly optimising the transforms and sets of scalar compression rates across all receivers has previously been established, it requires intensive numerical solutions, and is not a practical scalable solution for real time implementation.

A solution is proposed in which each receiver applies the KLT to its received signal, before compressing the transformed signal using a set of scalar compressors with optimised rate allocations. These rate allocations are jointly optimised across all receivers in order to maximise the sum capacity achieved under joint MMSE-SIC or MMSE detection. The solution to this optimisation has an iterative structure, but benefits from the availability of closed form expressions at each iteration, meaning – unlike prior work – numerical solvers are not required. The complexity is $\mathcal{O}(K^3)$ at each iteration, and a small number of iterations (3-6) are required for a good solution, with the optimisation performed once per coherence block. Furthermore, numerical results show that this approach suffers only a small performance penalty compared to the optimal transform coding scheme (which has significantly higher complexity).

The proposed scheme using transform coding with joint rate allocation represents a scalable solution for efficient distributed MIMO signal compression. Numerical examples demonstrate

that with 4 remote receivers, each equipped with 8 antennas and 1.5 Gbps fronthaul capacity, a total mean user throughput of over 420 Mbps could be simultaneously supplied to 8 users within a 100 MHz channel – an improvement of 100 Mbps per user compared to the case where the compression rates are not jointly optimised. As in the single receiver case, deploying additional antennas at each receiver is shown to provide array gain and improve energy efficiency. The importance of having good quality CSI is demonstrated.

The findings of this chapter point to some important areas for further development. Firstly, the analysis provided here is all based on the Gaussian scalar compression model, a theoretical model which – whilst widely used – only approximates the performance of practical entropy-coded quantization schemes. Investigating the complexity trade-offs of different practical compression schemes and validating the proposed solution using bit level simulations is therefore the next development step. Secondly, the results in this chapter have demonstrated that in the fronthaul-limited region transform coding utilises the available fronthaul capacity most effectively at high SNR. Whilst this is a reasonable operating assumption for dense distributed MIMO deployments, the higher transmit power brings diminishing performance returns in this region, reducing the energy efficiency of the system – this should be characterised to determine appropriate operating conditions & power control schemes.

Whilst discussed briefly in this chapter, a full investigation into the signalling overheads related to the transfer of CSI over fronthaul is required, assessing the impact of mobility and channel coherence times – in practice this will impact the fronthaul data throughput that can be achieved. Additionally, the impact of wider aspects of fronthaul signalling that are beyond the scope of this work should be considered, such as latency requirements and the impact of fronthaul outage (e.g. due to blockage of wireless links).

A particularly interesting insight provided by this chapter is that when there are an overall excess of antennas and the system operates with limited fronthaul capacity the jointly optimised rate allocations tend to be sparse – only a subset of the signal components at each receiver are compressed. This solution structure points towards an alternative approach to signal compression for distributed MIMO networks, based on *dimension reduction*: this is the subject of Chapter 5.

Chapter 5

Dimension Reduction for Distributed MIMO C-RAN

Recent research into large scale distributed MIMO systems has demonstrated the benefits of a cellular architecture in which the total number of BS antennas exceeds the number of users being actively served [152], [210]. However, in such systems the high dimensionality of the BS signals – both those received on the uplink, and those precoded for transmission on the downlink – becomes problematic when the distributed remote radio heads (RRH) are connected to the central processor (CP) via fronthaul connections with limited capacity [98].

Dimension reduction is a widely-used data processing strategy in which the underlying sparsity of a high dimensional signal is exploited to produce a low dimensional representation that can be more easily stored, transferred and processed [56]. This idea has recently been extended to cases where the high dimensional signal is distributed between a number of distinct nodes – such as in a network of sensors [188] – *distributed* dimension reduction.

Somewhat surprisingly, the explicit application of distributed dimension reduction to distributed MIMO networks has previously received very little research attention. Considering a MIMO C-RAN architecture in which multiple users are served by multiple geographically distributed multi-antenna remote radio heads (as in Section 4.4), this final research chapter investigates the use of dimension reduction-based signal compression for both the uplink and downlink signals.

Low complexity dimension reduction-based schemes for uplink signal compression and downlink signal precoding are proposed, taking closely related 'dual' forms:

- On the uplink, each remote receiver applies a linear dimension reduction filter to its multi-antenna received signal to produce a reduced dimension signal representation. This is then quantized and forwarded over fronthaul to the CP, as shown in Figure 5.1. The CP uses the ensemble of reduced dimension signals to jointly detect the user symbols.

- On the downlink, each remote transmitter receives a low dimension signal over fronthaul, which it then beamforms using a larger number of antennas. The signal precoding takes place in two stages: at the CP, an inner precoder takes the user symbols and produces the low dimensional signals, which are then quantized and transferred over fronthaul to the transmitter, which perform the outer precoding, as shown in Figure 5.2.

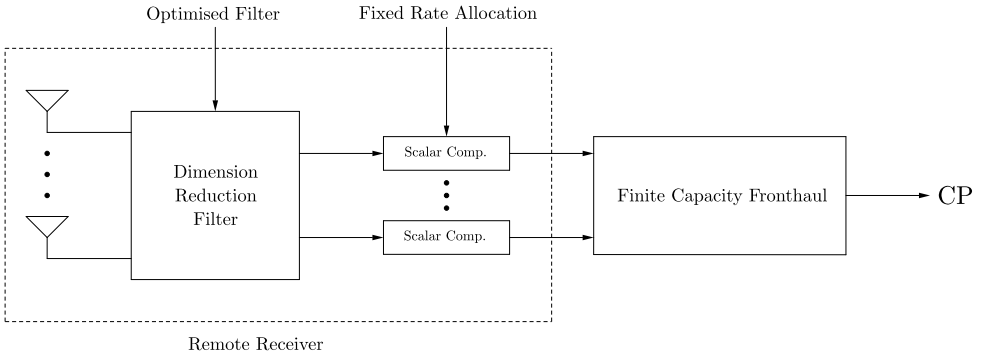

Figure 5.1: Block diagram of proposed dimension reduction uplink fronthaul compression scheme (single receiver shown).

The chapter begins by investigating the design of appropriate dimension reduction filters for the distributed MIMO uplink. Here, it is chosen to design the filters under a joint mutual information criteria, such that the information that the set of reduced dimension signals jointly provide about the user symbols is maximised. It is shown that *the optimal filter* for any given receiver is a *truncated version* of the *conditional KLT* (T-CKLT) – a transform found in other distributed dimension reduction applications [62] – where here the (statistical) conditioning is performed with respect to the other reduced dimension signals. Whilst the problem of finding the optimal set of dimension reduction filters is non-convex, a stationary point is shown to be found using a block coordinate ascent procedure where the filters are iteratively updated at each receiver in turn.

It is shown analytically that the proportion of information in the received signals that is lost due to dimension reduction decreases as SNR increases, whilst numerical examples show that it is possible to produce a good signal representation with a significantly reduced dimension. It is then shown using numerical examples that for a fixed signal dimension, increasing the number of antennas at each receiver is beneficial – providing both an array gain and decreasing the eigenvalue spread of the reduced dimension channel.

A second, simpler, scheme – the MF-GS scheme – is then proposed, where dimension reduction is achieved by matched filtering the received signal at each receiver using a subset of the available local user channel vectors. Though sub-optimal, the MF-GS filters incur a relatively small performance penalty relative to the T-CKLT filters, whilst providing benefits in terms of both computational complexity and signalling overheads. Both of the dimension reduction filter design methods are shown to be straightforward to adapt for the case of imperfect CSI.

The application of lossy signal quantization/compression to the reduced dimension signals is then investigated for the case where a set of equal resolution (equal rate) scalar quantizers are used. This combination of dimension reduction and scalar compression has clear parallels to the transform coding approach investigation in Section 4.4, with the distinction that here, good performance is achieved by exploiting the data reduction achieved from applying the jointly designed dimension reduction transforms, as opposed to from using a fixed transform and optimising the rate allocation.

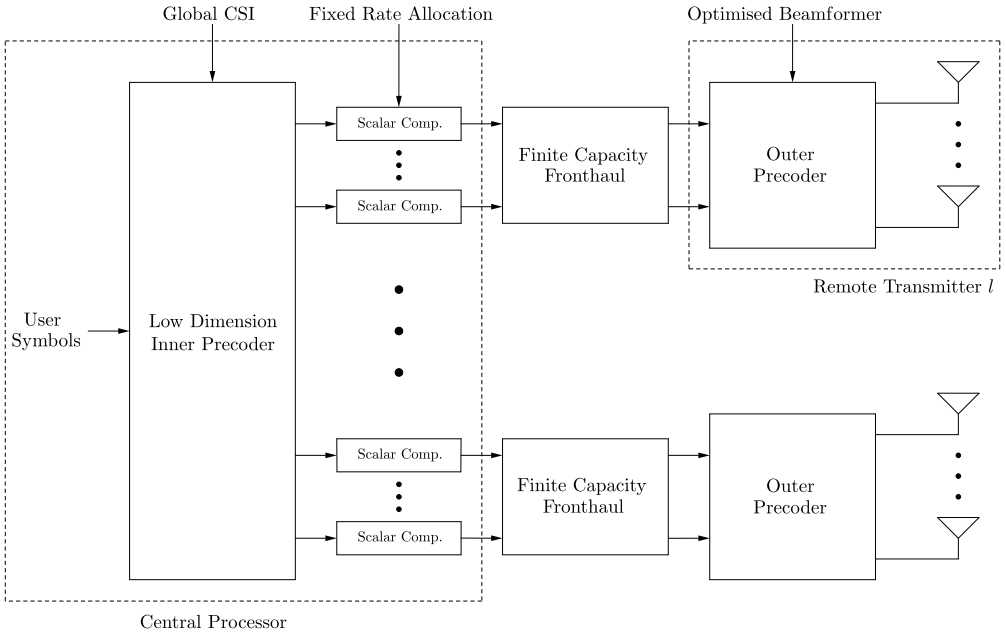

Figure 5.2: Block diagram of proposed two-stage fronthaul-aware downlink precoding scheme.

First, useful insights into the performance of reduced dimension compression are provided by analysing the MIMO sum capacity under Gaussian scalar compression and MMSE-SIC detection. A high SNR approximation is derived which shows that in the quantization-noise limited region the *sum capacity scales approximately linearly with the available fronthaul capacity*, with a gradient that is *inversely proportional to the signal dimension*. Since dense distributed MIMO deployments are expected to operate at high SNR, this provides a more rigorous justification for the intuition that reducing the signal dimension is an effective way of increasing fronthaul efficiency. Numerical results then show that, by choosing the optimal signal dimension for the given fronthaul capacity, the use of dimension reduction with equal rate scalar quantization can effectively match, and even outperform, the sum capacity achieved by either the P2P-RA or P2P-SCA schemes – it is a *quasi-optimal compression strategy*.

The use of T-CKLT or MF-GS dimension reduction filters in conjunction with simple *fixed-rate* scalar quantization and linear MMSE symbol detection is then studied, as a practical, scalable approach to signal compression that could be used in large distributed MIMO networks. Approximations of the user SINRs are derived, and validated using bit-level simulations. A numerical case study is then provided for a dense MIMO deployment, showing that, for example, under reasonable operating conditions, 16 users can be served a mean uplink throughput of 300 Mbps in a 60 MHz channel by 16 remote receivers each equipped with a 1 Gbps fronthaul connection.

Finally, the ideas from uplink dimension reduction are adapted for a two-stage fronthaul-aware downlink precoding scheme – establishing dimension reduction as a bidirectional approach. Exploiting the duality between the MIMO uplink and downlink, it is argued that the T-CKLT and MF-GS methods can be used to design effective outer precoders that each transmitter can use to transmit a low dimensional signal using a larger number of antennas. Having

chosen and fixed these outer precoders, inner zero-forcing symbol precoders are then designed by considering the effective channel provided by the concatenation of the outer precoders and MIMO propagation channels.

For downlink signal compression, equal resolution fixed-rate scalar quantization is applied to the low dimension signals at the output of the inner precoder. This has the effect of introducing quantization noise into the precoded signals, which is then transmitted towards the users, degrading their performance, as captured by the derived SINR expressions. The quantization noise characteristics depend on the user power allocations, and so to address the fact that each user is impacted differently by quantization noise, a max-min power control scheme is proposed to balance the user performance whilst ensuring per-transmitter power constraints are met.

Numerical results are again provided for a dense MIMO deployment and 60 MHz channel, showing that under realistic power constraints, user mean downlink throughputs on the order of 300 Mbps can be provided by 16 transmitters each operating with 1 Gbps fronthaul capacity, with a clear benefit over conventional precoding demonstrated.

5.1 Chapter Overview

The chapter has the following general structure:

- Section 5.2 provides background to the use of dimension reduction in MIMO C-RAN systems. It begins by briefly outlining the theory behind statistical dimension reduction and compressed sensing approaches, before reviewing previous applications to distributed MIMO networks with limited fronthaul capacity.

- Section 5.3 considers the design of dimension reduction filters for the distributed MIMO uplink, and analyses their performance. First, the filters are optimised according to joint mutual information criteria. The performance of the reduced dimension MIMO channel produced is then studied, using both mathematical analysis and numerical simulations. An alternative filter design scheme based on matched filtering is designed, and its potential for reducing signalling overheads is analysed before, finally, an adaptation of the dimension reduction techniques for channels with imperfect CSI is outlined.

- Section 5.4 studies the use of dimension reduction in conjunction with lossy signal compression. It begins by finding new dimension reduction filters that are optimised under a sum capacity objective that includes the impact of quantization noise under Gaussian scalar compression. The sum capacity scaling at high SNR is then analysed, and performance compared to other fronthaul compression schemes. Finally, the performance of a practical low complexity dimension reduction-based compression scheme that uses fixed-rate scalar quantizers and linear detection is studied using numerical examples.

- Section 5.5 applies the uplink signal compression ideas to the distributed MIMO downlink. It begins by outlining a two stage precoding scheme, where, exploiting a duality with the uplink, the outer precoders are designed by reversing the dimension reduction filters designed in Section 5.3, before an inner precoder is designed under a zero-forcing criteria.

It is then proposed to use max-min power control to account for the effects of quantization noise, with numerical results presented.

5.1.1 Novel Contributions

The key contributions to the state-of-the-art made in this chapter are:

- **Showing that the dimension reduction filters that maximise the joint mutual information between the set of reduced dimension uplink signals and user symbols are a truncated form of the conditional KLT, with a stationary point found using block coordinate ascent, Section 5.3.1.** To the author's knowledge, distributed dimension reduction under a joint mutual information criteria has not previously been studied (in a communications context or possibly beyond), and hence this result could also have wider applicability, e.g. in sensor networks.

- **Numerical results demonstrating that joint signal representations with significantly reduced dimensionality can be found that accurately preserve the key characteristics of the full dimension signal, Section 5.3.2.** For example, in a system with 8 users, 4 remote receivers and 8 antennas per receiver, a representation using 3 or 4 dimensions per receiver is sufficient to produce a reduced dimension MIMO channel with similar eigenvalue spread to the full channel, incurring negligible loss of information.

- **An alternative dimension reduction filtering scheme based on matched filtering, that can be used for significantly reducing signalling overheads between the receivers and CP, Section 5.3.3.** In this scheme, the dimension reduction filters are constructed using a defined subset of the local user channel vectors. Assuming CSI is initially obtained at the receivers, dimension reduction can be directly applied and only the reduced dimension channel matrices transferred over fronthaul to the CP.

- **A high SNR approximation showing that, under Gaussian scalar compression, the MIMO sum capacity in the quantization/fronthaul-limited region scales approximately linearly with the available fronthaul capacity, and inversely with the signal dimension, Section 5.4.1.** This approximation is found by taking lower and upper capacity bounds at the high SNR asymptotic limit, and provides a theoretical rationale for the use of dimension reduction-based signal compression.

- **Numerical results demonstrating that the use of dimension reduction with scalar compression can be a highly efficient fronthaul compression strategy, Section 5.4.1.** Comparisons with the SCA-P2P & SCA-RA transform coding fronthaul compression schemes from the previous chapter show that dimension reduction-based fronthaul compression with optimal signal dimension selection can outperform both schemes in terms of sum capacity/fronthaul efficiency.

- **A practical distributed MIMO fronthaul compression and signal detection scheme based on the use of dimension reduction filtering, simple fixed-rate scalar quantizers and linear symbol detection, Section 5.4.2, Section 5.4.3.**

SINR expressions are derived and validated by bit-level simulation, with example throughput figures for a case study dense deployment operating with imperfect CSI provided.

- **A downlink 'dual' of the uplink reduced dimension compression scheme, in which two-stage precoding is used to reduce the dimensionality of the signals transferred over fronthaul, Section 5.5.1.** Exploiting a simple uplink/downlink duality, it is shown that outer precoders – applied at the receivers – can be designed using the same methods as the dimension reduction filters, but applied in reverse to take a low dimensional signal – precoded by an inner precoder at the CP – and beamform it using a larger number of antennas.

- **A downlink max-min power allocation scheme that optimises user power allocations to mitigate the performance degradation caused by applying quantization to the inner-precoded signals before tranferring them over fronthaul, Section 5.5.2.** This is shown to be able to significantly improve the capacities of the worst users when operating in the quantization-limited region.

5.1.2 Published Work

A matched-filtering based signal compression scheme was published in [221]. This paper mainly contains the content of Section 5.3.3, but uses local transform coding applied at each receiver for lossy compression, rather than the direct scalar compression considered in this chapter. The paper includes the adaptation for channels with imperfect CSI and capacity expressions for both MMSE-SIC and MMSE detection.

As with the other research chapters, key results in this chapter remain unpublished. Most significantly, the T-CKLT dimension reduction filters, high SNR sum capacity scaling and downlink two-stage precoding scheme have not been published. It is planned to write a journal paper that brings together the uplink and downlink scheme to show the benefits of reduced dimension signalling in C-RAN networks with limited fronthaul capacity.

5.2 Background

The potential benefits of deploying an overall excess of BS antennas in a distributed MIMO C-RAN system are well known; increasing the number of antennas at each RRH provides diversity and array gain, whilst increasing the number of RRHs improves uniformity of coverage through macro-diversity. However, this growing signal dimensionality becomes an issue – on both the uplink and downlink – when the C-RAN fronthaul network has limited capacity, as discussed in Chapter 4.

Whilst the signals being received (or transmitted) by each of the ML antennas are distinct, all are linear combinations of the same K user uplink (or downlink) symbols. As the overall excess of BS antennas, ML/K, grows, the signals therefore become characterised by their sparsity – they can be represented in a signal basis where most of the coefficients are zero. Exploiting this sparsity to find reduced dimension representations of the transmit and receive

signals provides a first-order reduction in the quantity of data being transferred over fronthaul, and can potentially be used as a part of an efficient fronthaul compression scheme.

In the previous chapter it was shown that for a single remote radio head the sparse signal basis is given by the eigenvectors of the channel matrix, $\mathbf{U}$. On the uplink, projecting the received signal into the data signal subspace using the KLT reduces the signal dimension from M to K without loss of information, and is the first stage of the optimal transform coding compression scheme.

In the multi-RRH setting, producing a minimum dimension representation of the signal for signal compression is less straightforward, since each RRH only has access to a fraction of the received signal. The application of the local KLT at each receiver can exploit any local sparsity that exists when an individual receiver has an excess of antennas, $M > K$, but cannot exploit the *joint* sparsity that exists between the ensemble of received signals. In Chapter 4, this joint sparsity was instead implicitly accounted for by a joint rate allocation scheme that tended to encode a subset of the available signal components at each receiver.

This section gives some background to the problem of *distributed dimension reduction*, discussing some general approaches before turning attention to previous use of dimension reduction in MIMO C-RAN systems.

5.2.1 Distributed Dimension Reduction

The so-called 'curse-of-dimensionality' is frequently encountered across a wide range of signal processing applications, creating issues of scalability around signal storage, processing and transmission. Recently, there has been a growing interest in applications where those signals are distributed between a network of nodes, such as in sensor networks. This has led to the development of a number of different dimension reduction techniques that aim to exploit sparsity to reduce the signal dimensionality at each node whilst preserving the salient features of the signal ensemble.

Two approaches to distributed dimension reduction are now briefly discussed:

- Those that use knowledge of the signal statistics, e.g. signal correlations, to calculate reduced dimension signal approximations.
- Those based on compressive sensing, that do not use knowledge of the signal statistics when performing dimension reduction.

Statistical Methods

In the classical 'centralised' principal component analysis method, the best N-dimensional representation (in a mean squared error sense) of a vector is found by projecting it onto the N principal eigenvectors of the signal covariance matrix [92]. This is precisely the KLT as outlined in the previous chapter, and has found wide use beyond transform coding – for example in a communications context it is used in [100] to reduce the complexity of space-time processing in a single-user MIMO system.

The idea of principal component analysis was extended to the distributed setting in [62], which considers the scenario where each node observes a different part of a correlated vector, to

which it applies a dimension reducing transform. They investigate a number of problems relating to finding the reduced dimension representations that permit the minimum mean squared error approximation of the original vector.

It is shown when all but one node supplies its full observation that the optimal transform for the remaining node is given by the N principal eigenvectors of the conditional variance of its own signal, i.e. the basis in which the signal has the greatest variance when the correlated observations are provided by the other nodes. This is known as the *conditional* KLT (CKLT). For the case where all nodes apply dimension reduction a stationary point to the problem is found iteratively by updating the transform at each node, for which a slightly more involved closed form expression is provided.

A related problem is addressed in [188], but where each node is a sensor and the aim is not to reconstruct the set of sensor observations but to linearly estimate some other correlated quantity from reduced dimension representations of the observations. This has clear parallels to the distributed MIMO C-RAN model, which aims to jointly estimate the user symbols from the set of received signals. The optimal transforms here are also found in a round-robin iterative fashion, using somewhat tedious closed-form linear expressions that depend on the transforms at the other nodes and various cross-variance and covariance matrices.

With both of these methods the optimal transforms depend on the signals statistics at all nodes. The transforms must therefore be computed at some node within the network that has access to full statistical information, and then transferred to the other nodes.

Compressed Sensing

The compressed sensing (CS) approach is able to exploit sparsity in a signal to reduce the signal dimension without requiring any a priori statistical knowledge. The key idea is that when a full dimension signal vector is known to have a sparse representation in a certain basis (the 'sparse basis'), then any reduced dimension representation of the signal will preserve most of the information in the original providing the reduced signal dimension is sufficiently large compared to the number of non-zero elements in the sparse representation, and providing that all potential vectors that are sparse in the sparse basis have dense representations in the sensing (reduced dimension) signal basis [29]. This second condition is fulfilled by choosing the sensing basis to be 'incoherent' with the sparse basis. For example any signal that is known to be sparse in the frequency domain (i.e contain a small number of tones), will have a dense representation in the time domain, since the two bases are incoherent, and a random selection of time samples suffices for sparse signal recovery [27].

The 'restricted isometry property' (RIP) is often used to study the suitability of a sensing matrix for CS [28]. With high probability, a randomly chosen sensing basis is likely to be incoherent with any sparse basis, and obey the RIP [29], providing the dimension of the basis is sufficiently large – typically ~ 4 times the number of non-zero elements. Recovering the sparse signal then requires numerically solving a least squares problem that includes a ℓ_1 regularisation term that heavily favours sparse solutions.

The CS approach has been extended to distributed settings in [51] and [10] under a variety of different jointly sparse signal models.

The advantage of the CS approach over statistics based approaches is that a reduced dimension signal can be formed using random projections, with no knowledge of the signal statistics required. However, the sparse basis must still be known for signal recovery, and a larger signal dimension is generally required compared to the statistical approach.

5.2.2 Distributed Dimension Reduction in MIMO C-RAN

Distributed dimension reduction has been applied in a variety of forms to help reduce fronthaul loads in both uplink and downlink MIMO C-RAN. On the uplink these have mainly focused on compressed sensing techniques, whilst on the downlink the sparse beamforming approach has received attention.

MIMO C-RAN Uplink

On the uplink, various efforts have considered an architecture where the numbers of BS antennas in the network, ML, exceeds the number of active users, K.

Notably, in [175] distributed CS is applied to a network with single antennas RRHs by applying a sensing matrix across many subcarriers to produce reduced dimension observations at each RRH, which are then quantized and transferred across fronthaul. It is assumed that the channels to all users are known, but that only an unknown subset of the users are actively transmitting. Using the reduced dimension signals, the CP first jointly detects the active users and establishes rough symbol estimates using CS methods, and then performs conventional ZF symbol detection once the set of active users is known. The RIP is characterised for this case, and numerical results presented that show the active users can be accurately indentified and their symbols recovered. A related scheme is introduced in [233] that also performs user channel estimation, and in [120] that mitigates sources of narrowband interference.

The CS approach has also been applied to networks with multi-antenna RRHs. In [245] the results from [175] are adapted for the case where the sensing matrix is applied across the multi-antenna signal, rather than across different subcarriers. The RIP is studied, and numerical results presented shown for various configurations that indicate an overall signal dimension of at least $\sim 4K$ is required to recover the K user symbols.

It should be noted that the benefits of using CS in the above methods is for their ability to determine which users are actively transmitting – which could be relevant for example in a large sensor network. If the users are scheduled to transmit in advance then CS methods need not be directly applied.

In [128] a dimension reduction fronthaul compression scheme is proposed for multi-antenna C-RAN with known active users. In this scheme each RRH creates a reduced dimension signal by taking simple unweighted sums of the signals received at different antennas in different symbol time slots or on different subcarriers (i.e. by applying a binary linear transform with entries that are all either 0 or 1). The signal dimensionality is chosen by controlling how many received signal measurements to combine into each dimension, and the linear combinations taken by each receiver to not depend in any way on the signal statistics/channel realisation. Since combining signals from orthogonal resource blocks into the same signals creates additional inter-user performance, a parallel interference cancellation detection scheme is introduced to improve

performance. The scheme is shown to give good compression, particularly at low fronthaul rates, and significantly outperforms simple sample & forward compression.

Dimension reduction techniques have also been applied less directly. For example [118] uses analogue beamforming to reduce the signal dimension prior to sampling and compression, whilst, similarly to the findings in Section 4.4.3, the joint rate allocation in [123] was found to tend to implicitly perform dimension reduction on the signal at reduced fronthaul capacities.

MIMO C-RAN Downlink

On the uplink of the MIMO C-RAN, each RRH receives the uplink transmissions of all users within its coverage area. On the downlink, however, each RRH can potentially transmit precoded data symbols to only a *subset* of users, whilst the network as a whole, by coordination of the different RRHs, ensures every user is served. This opens up new opportunities for reducing the downlink fronthaul traffic load.

Under linear downlink precoding, the precoding weights can equally be applied to the downlink symbols centrally at the CP, or locally at the RRHs. When performed locally, the RRH must have access to the downlink symbols for the users it is serving, and its required fronthaul capacity is therefore equal to the sum capacities of those users (plus any overheads). The 'sparse beamforming' approach exploits this for fronthaul load reduction, by designing joint precoding strategies that require each RRH to only serve a subset of the users in its coverage area [48]. When the number of antennas at each RRH is greater than the number of users served by the RRH, this can be interpreted as a dimension reduction strategy in the sense that the dimensionality of the signal transferred over fronthaul is lower than the dimensionality of the transmit signal.

A variety of sparse beamforming strategies have been proposed, e.g [90], [126] & [194], generally making use of sparse optimisation tools to determine the user-RRH association. The method in [48] explicitly incorporates a fronthaul constraint in the problem formulation, and uses a successive convex approximation approach to jointly determine the precoding matrices and user associations.

An alternative fronthaul compression strategy is to perform all symbol precoding centrally at the CP, and then compress the precoded signals for transfer over fronthaul to the RRHs. Dimension reduction is exploited in this context in [110], where analogue beamforming is used in conjunction with low-dimension digital beamforming to reduce the dimensions of the signals transferred over fronthaul.

5.3 Reduced Dimension Distributed MIMO Uplink Channels

As the excess of antennas in a finite fronthaul capacity distributed MIMO system grows, it becomes desirable to find reduced dimension signal representations in order to reduce the quantity of data transferred over fronthaul.

On the uplink, the ensemble of signals at the L receivers can be expressed as a global received signal,

$$\mathbf{y}_{\mathrm{G}} = \mathbf{H}_{\mathrm{G}}\mathbf{x} + \boldsymbol{\eta}. \tag{5.1}$$

From the discussion in Section 4.3.2, the portion of this signal that contains useful information about the user symbols can be fully represented by a K-dimensional signal by applying, for example, the KLT to $\mathbf{y}_{\mathrm{G}}$.

However, this signal representation cannot be produced in a distributed setting where each receiver only has access to its own received signal. When each receiver is equipped with $M > 1$ antennas, *distributed dimension reduction* can instead be performed by each receiver locally applying a linear dimension reduction filter, $\mathbf{A}_l \in \mathbb{C}^{M \times N}$ to its received signal to produce a signal with $N < M$ dimensions[1],

$$\mathbf{z}_l = \mathbf{A}_l^{\dagger} \mathbf{y}_l. \tag{5.2}$$

At a network level this can be expressed as an NL-dimensional signal,

$$\mathbf{z}_{\mathrm{G}} = \mathbf{A}_{\mathrm{G}}^{\dagger} \mathbf{y}_{\mathrm{G}}, \tag{5.3}$$

where $\mathbf{z}_{\mathrm{G}} \in \mathbb{C}^{NL}$ is the ensemble of L reduced dimension signals and $\mathbf{A}_{\mathrm{G}} \in \mathbb{C}^{ML \times NL}$ is the equivalent block-diagonal dimension reduction filter,

$$\mathbf{z}_{\mathrm{G}} = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_L \end{bmatrix}, \qquad \mathbf{A}_{\mathrm{G}} = \begin{bmatrix} \mathbf{A}_1 & & \\ & \ddots & \\ & & \mathbf{A}_L \end{bmatrix}. \tag{5.4}$$

At each receiver, the user symbols are contained within a t-dimensional signal space, where $t = \min(M, K)$, meaning that – unlike with the global KLT dimension reduction – any joint representation of the received signals that uses less than Lt dimensions in total necessarily involves *information loss*.

This section addresses the task of finding the filters that produce a 'good' joint representation of the distributed MIMO received signals using $N < t$ dimensions at each receiver. Specifically, the dimension reduction filters are chosen to provide the maximum information about the user symbols, by maximising the joint mutual information,

$$\mathcal{I}\big(\mathbf{z}_1, \ldots, \mathbf{z}_L; \mathbf{x}\big) = \mathcal{I}\big(\mathbf{A}_1^{\dagger}\mathbf{y}_1, \ldots, \mathbf{A}_L^{\dagger}\mathbf{y}_L; \mathbf{x}\big). \tag{5.5}$$

A procedure for finding a stationary point to this problem is provided, before a reduced complexity method based on matched filtering is developed.

5.3.1 Maximum Mutual Information Distributed Dimension Reduction

The reduced dimension signal at receiver l is given by

$$\mathbf{z}_l = \mathbf{A}_l^{\dagger} \mathbf{H}_l \mathbf{x} + \boldsymbol{\eta}_{\mathrm{A}} \tag{5.6}$$

where

$$\mathbb{E}\big[\boldsymbol{\eta}_{\mathrm{A}} \boldsymbol{\eta}_{\mathrm{A}}^{\dagger}\big] = \mathbf{A}_l^{\dagger} \mathbf{A}_l. \tag{5.7}$$

[1]It is assumed here for simplicity that each receiver uses the same signal dimension. Many of the ideas can be immediately generalised to the case where different signal dimensions are used.

Without loss of generality, attention may be restricted to semi-orthogonal filters,

$$\mathbf{A}_l^\dagger \mathbf{A}_l = \mathbf{I}_N, \tag{5.8}$$

since, using the QR decomposition, any non-orthogonal filter with linearly independent columns, $\tilde{\mathbf{A}}_l \in \mathbb{C}^{M\times N}$, can be written

$$\tilde{\mathbf{A}}_l = \mathbf{A}_l \mathbf{T}_l, \tag{5.9}$$

where $\mathbf{T}_l \in \mathbb{C}^{N\times N}$ is invertible and hence by the data processing inequality [47] does not affect the mutual information. The filtering operation thus amounts to taking the projection of the received signals onto a subspace which has the columns of $\mathbf{A}_l$ as a basis.

Each reduced dimension signal can be expressed as the output of an equivalent reduced dimension MIMO channel,

$$\mathbf{z}_l = \mathbf{G}_l \mathbf{x} + \boldsymbol{\eta}, \tag{5.10}$$

and joint mutual information is therefore

$$\mathcal{I}\big(\mathbf{z}_1, \ldots, \mathbf{z}_L; \mathbf{x}\big) = \log_2 \det\big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \mathbf{G}_l^\dagger \mathbf{G}_l\big) \tag{5.11}$$

$$= \log_2 \det\big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \mathbf{H}_l^\dagger \mathbf{A}_l \mathbf{A}_l^\dagger \mathbf{H}_l\big). \tag{5.12}$$

Finding a globally optimal set of dimension reduction filters is infeasible, since the optimisation

$$\begin{aligned} \underset{\mathbf{A}_1,\ldots,\mathbf{A}_L}{\text{maximise}} \quad & \log_2 \det\big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \mathbf{H}_l^\dagger \mathbf{A}_l \mathbf{A}_l^\dagger \mathbf{H}_l\big) \\ \text{subject to} \quad & \mathbf{A}_l^\dagger \mathbf{A}_l = \mathbf{I}_N, \quad \forall l, \end{aligned} \tag{5.13}$$

is non-convex. However, a block coordinate ascent (BCA) approach can be used to find a stationary point, by maximising in turn over a single $\mathbf{A}_l$ whilst holding the others constant. This is based on the mutual information expansion

$$\mathcal{I}(\mathbf{z}_1, \ldots, \mathbf{z}_L; \mathbf{x}) = \mathcal{I}(\mathbf{z}_l; \mathbf{x} | \mathbf{z}_l^{\mathsf{c}}) + \mathcal{I}(\mathbf{z}_l^{\mathsf{c}}; \mathbf{x}), \tag{5.14}$$

where $\mathbf{z}_l^{\mathsf{c}} = \{\mathbf{z}_1 \ldots \mathbf{z}_{l-1}, \mathbf{z}_{l+1} \ldots \mathbf{z}_L\}$ is the ensemble of reduced dimension signals from all receivers except receiver l. Only the first term in (5.14) is dependent on $\mathbf{A}_l$, and can be expanded

$$\mathcal{I}(\mathbf{z}_l; \mathbf{x} | \mathbf{z}_l^{\mathsf{c}}) = \mathcal{H}\big(\mathbf{z}_l | \mathbf{z}_l^{\mathsf{c}}\big) - \mathcal{H}\big(\mathbf{z}_l | \mathbf{x}, \mathbf{z}_l^{\mathsf{c}}\big) \tag{5.15}$$

$$= \mathcal{H}\big(\mathbf{z}_l | \mathbf{z}_l^{\mathsf{c}}\big) - \mathcal{H}\big(\boldsymbol{\eta}_{\text{A}}\big) \tag{5.16}$$

The second term here is independent of $\mathbf{A}_l$, and the optimal transform at receiver l therefore maximises the conditional entropy of $\mathbf{z}_l$ given $\mathbf{z}_l^{\mathsf{c}}$,

$$\mathcal{H}\big(\mathbf{z}_l | \mathbf{z}_l^{\mathsf{c}}\big) = \log_2 \det\big(\mathbf{I}_N + \rho \mathbf{A}_l^\dagger \mathbf{H}_l \mathbf{Q}_l \mathbf{H}_l^\dagger \mathbf{A}_l\big) + \log_2 (2\pi e)^N \tag{5.17}$$

where (see Appendix 3.1)

$$\mathbf{Q}_l = \left(\mathbf{I}_K + \rho \sum_{i \neq l} \mathbf{H}_i^\dagger \mathbf{A}_i \mathbf{A}_i^\dagger \mathbf{H}_i\right)^{-1}. \tag{5.18}$$

The optimal dimension reduction filter at receiver l is then the solution to

$$\begin{aligned} \underset{\mathbf{A}_l}{\text{maximize}} \quad & \det\left(\mathbf{I}_N + \rho \mathbf{A}_l^\dagger \mathbf{H}_l \mathbf{Q}_l \mathbf{H}_l^\dagger \mathbf{A}_l\right) \\ \text{subject to} \quad & \mathbf{A}_l^\dagger \mathbf{A}_l = \mathbf{I}_N. \end{aligned} \tag{5.19}$$

Using Poincare's separation theorem [14] it can be shown (see Appendix 3.2) that the optimal filter corresponds to the N principal eigenvectors of $\mathbf{H}_l \mathbf{Q}_l \mathbf{H}_l^\dagger$ (those corresponding to the maximum eigenvalues). These are the eigenvalues of the conditional variance of $\mathbf{y}_l$ given $\mathbf{z}_l^{\mathsf{c}}$,

$$\mathbb{E}\left[\mathbf{y}_l \mathbf{y}_l^\dagger \middle| \mathbf{z}_l^{\mathsf{c}}\right] = \mathbf{I}_M + \rho \mathbf{H}_l \mathbf{Q}_l \mathbf{H}_l^\dagger, \tag{5.20}$$

and hence the optimal transform is a truncated form of the conditional KLT (T-CKLT herein) as outlined in [62].

A stationary point to (5.13) may accordingly be found using a BCA procedure [162], iteratively updating the $\mathbf{A}_l$ in turn, as shown in Algorithm 4. Here the $\mathbf{A}_l$ are initialised using the first N outputs of the local KLT filters[2].

Algorithm 4 T-CKLT block-coordinate ascent (BCA) dimension reduction filter design

inputs: $\mathbf{H}_l \quad \forall l$
$\mathbf{A}_l \leftarrow N$ principal eigenvectors of $\mathbf{H}_l \mathbf{H}_l^\dagger \quad \forall l$
for $j = 1 : j_{\max}$ **do**
  for $l = 1 : L$ **do**
    $\mathbf{Q}_l \leftarrow \left(\mathbf{I}_K + \rho \sum_{i \neq l} \mathbf{H}_i^\dagger \mathbf{A}_i \mathbf{A}_i^\dagger \mathbf{H}_i\right)^{-1}$
    $\mathbf{A}_l \leftarrow N$ principal eigenvectors of $\mathbf{H}_l \mathbf{Q}_l \mathbf{H}_l^\dagger$
  end for
end for
outputs: $\mathbf{A}_l \quad \forall l$

At each sub-iteration, $\mathcal{I}(\mathbf{z}_1, \ldots, \mathbf{z}_L; \mathbf{x})$ monotonically increases, and hence Algorithm 4 converges to a stationary point of (5.13).

The main computations required at each sub-iteration are:

- Inversion of $\mathbf{Q}_l$, complexity $\mathcal{O}(K^3)$.
- Calculation of $\mathbf{H}_l \mathbf{Q}_l \mathbf{H}_l^\dagger$, complexity $\mathcal{O}(MK^2)$.
- Singular value decomposition $\mathcal{O}(M^3)$.

Assuming a deployment with $K > M$, the algorithm has computational complexity $\mathcal{O}(j_{\max} K^3 L)$. Simulations indicate that, when initialised using the KLT, a small number of iterations ($j_{\max} \leq$

[2]The local KLT can easily be shown to be the optimal dimension reduction filter in a single receiver setting by setting $L = 1$, which leads to $\mathbf{Q}_l = \mathbf{I}_K$.

3) are typically required to converge to within a practical tolerance of the maximum. This is illustrated in Figure 5.3, for two different system sizes and the simulation configuration described in Section 2.5.2. More extensive simulations (not shown) indicate that $j_{\max}$ does not scale with L, and hence the computational complexity of using the T-CKLT BCA method for finding the filters scales linearly with the number of receivers deployed. Whilst convergence to a stationary point, rather than global optimum, of the original problem is achieved, further simulations (also not shown) using different initialisations suggest that the variation between local maxima is not significant.

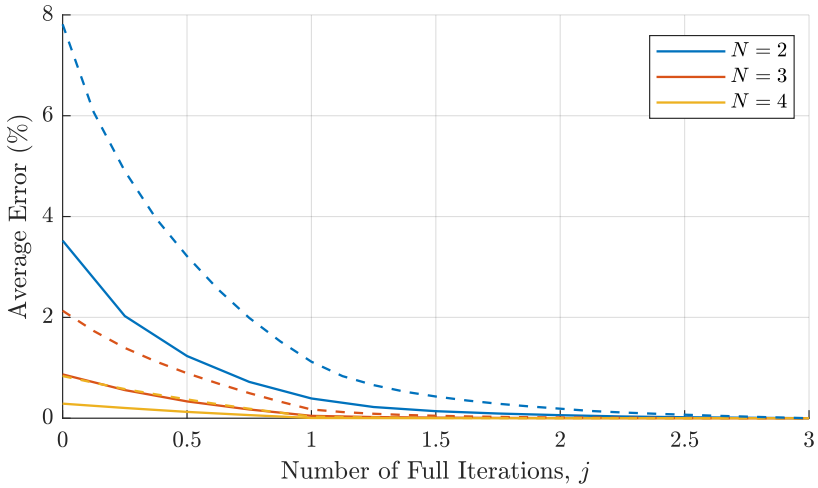

Figure 5.3: Convergence of T-CKLT BCA algorithm, $\rho = 15$ dB, $M = 8$. Solid line: $K = 8, L = 4$, dashed line: $K = 16, L = 8$.

The filters rely on global CSI, and must therefore be calculated at the CP and fed back to the receivers, incurring fronthaul overheads (see Section 5.3.3).

Dimension Reduction at High SNR

Rewriting the channel matrix to include user power scalings, $\mathbf{H}_l = \bar{\mathbf{H}}_l \mathbf{P}_l$, at high SNR (providing the inverse exists)

$$\bar{\mathbf{H}}_l \mathbf{P}^{1/2} \mathbf{Q}_l \mathbf{P}^{1/2} \bar{\mathbf{H}}_l^\dagger \approx \frac{1}{\rho} \bar{\mathbf{H}}_l \mathbf{P}^{1/2} \Big(\sum_{i \neq l} \mathbf{P}^{1/2} \bar{\mathbf{H}}_i^\dagger \mathbf{A}_i \mathbf{A}_i^\dagger \bar{\mathbf{H}}_i \mathbf{P}^{1/2} \Big)^{-1} \mathbf{P}^{1/2} \mathbf{H}_l^\dagger \tag{5.21}$$

$$= \frac{1}{\rho} \bar{\mathbf{H}}_l \Big(\sum_{i \neq l} \bar{\mathbf{H}}_i^\dagger \mathbf{A}_i \mathbf{A}_i^\dagger \bar{\mathbf{H}}_i \Big)^{-1} \bar{\mathbf{H}}_l^\dagger \tag{5.22}$$

and the receive filters are independent of the individual user transmit powers. This is a useful result since it implies that at high SNR the receive filters may be chosen independently of the user power control coefficients.

5.3.2 Reduced Dimension Channels

With a reduced signal dimension $N < t$ the signal representations produced by the T-CKLT BCA procedure can only approximate the original received signals. Fundamentally, for dimension reduction to be of any use in distributed MIMO networks, the dimension reduction filters must be able to capture a significant proportion of the information present in the full received signal in a lower dimensional signal representation. If not, then the dimensionality of the data can more simply be reduced by deploying fewer antennas at each receiver, and the use of dimension reduction techniques is unnecessary. This section therefore considers two questions:

- How well do the reduced dimension signals approximate the original received signals & how many signal dimensions are required for a good approximation?
- With a fixed signal dimension at each receiver are there any benefits to increasing the number of antennas deployed?

Clearly, finding complete answers to these questions involves mathematical analysis with specific channel models that, due to the complicated structure of the distributed MIMO user channels, is challenging and beyond the scope of this work. However, some general insights can be found using simple analysis and numerical examples.

In analysing the performance, parallels with the well known MIMO antenna selection problem are noted, and following a similar method to [71] the total information lost due to dimension reduction can be considered

$$\mathcal{L} = \mathcal{I}\big(\mathbf{y}_l, \ldots, \mathbf{y}_L; \mathbf{x} | \mathbf{z}_1, \ldots, \mathbf{z}_L\big). \tag{5.23}$$

Defining a matrix $\overline{\mathbf{A}}_l \in \mathbb{C}^{M\times(M-N)}$ that spans the complementary subspace to $\mathbf{A}_l$, i.e. $\mathbf{A}_l^\dagger \overline{\mathbf{A}}_l = \mathbf{0}$, the signal component discarded at receiver l during dimension reduction is $\overline{\mathbf{A}}_l^\dagger \mathbf{y}_l$ and the corresponding equivalent channel is $\overline{\mathbf{G}}_l = \overline{\mathbf{A}}_l^\dagger \mathbf{H}_l$. Similarly to [71], it can then be shown that

$$\mathcal{L} = \log_2 \det\Big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \overline{\mathbf{G}}_l^\dagger \overline{\mathbf{G}}_l \Big(\mathbf{I}_K + \rho \sum_{i=1}^{L} \mathbf{G}_i^\dagger \mathbf{G}_i\Big)^{-1}\Big). \tag{5.24}$$

This is monotonic in ρ, and can be upper bounded (where the inverse exists)

$$\mathcal{L} \leq \lim_{\rho\to\infty} \log_2 \det\Big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \overline{\mathbf{G}}_l^\dagger \overline{\mathbf{G}}_l \Big(\mathbf{I}_K + \rho \sum_{i=1}^{L} \mathbf{G}_i^\dagger \mathbf{G}_i\Big)^{-1}\Big) \tag{5.25}$$

$$= \log_2 \det\Big(\mathbf{I}_K + \sum_{l=1}^{L} \overline{\mathbf{G}}_l^\dagger \overline{\mathbf{G}}_l \Big(\sum_{i=1}^{L} \mathbf{G}_i^\dagger \mathbf{G}_i\Big)^{-1}\Big). \tag{5.26}$$

This bound is independent of ρ, implying that at high SNR dimension reduction causes a constant performance loss that depends only on the channel and reduced dimension filters. In contrast, the joint mutual information increases with ρ,

$$\mathcal{I}\big(\mathbf{z}_1, \ldots, \mathbf{z}_L; \mathbf{x}\big) > \log_2 \rho + \log_2 \det\Big(\sum_{l=1}^{L} \mathbf{G}_l^\dagger \mathbf{G}_l\Big), \tag{5.27}$$

and as a result the proportion of information lost due to dimension reduction vanishes as $\rho \to \infty$, as shown in Figure 5.4 for a system with $K = 8$, $M = 8$ & $L = 4$. This scaling holds for any $\mathbf{A}_l$ at high SNR, but $\mathcal{L}$ is minimised (to a stationary point) by the T-CKLT BCA method. Figure 5.4 shows the results for a configuration with 8 users and 4 receivers each with 8 antennas (i.e. $t = 8$), all randomly distributed within a 200 m $\times$ 200 m area – representing a dense urban deployment. The T-CKLT method is able to represent almost all of the received information for signal dimension $N \geq 2$. This contrasts with the use of random filters, as used in the compressed sensing approaches, which incur considerable information loss.

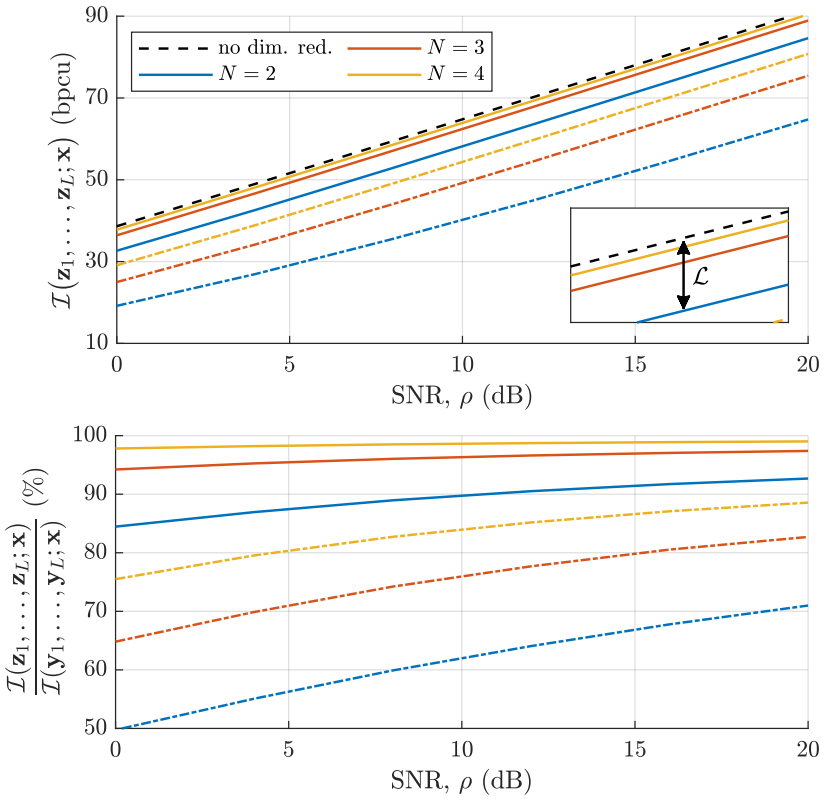

Figure 5.4: Reduced dimension mutual information scaling with ρ, $K = 8$, $M = 8$, $L = 4$ ($t = 8$). Solid line: T-CKLT filters, dot-dash line: random semi-orthogonal filter.

The reduced dimension channel has the eigendecomposition,

$$\sum_{l=1}^{L} \mathbf{G}_l \mathbf{G}_l^\dagger = \mathbf{\Theta} \mathbf{\Gamma} \mathbf{\Theta}^\dagger \tag{5.28}$$

where $\mathbf{\Theta} \in \mathbb{C}^{K \times K}$ are eigenvectors, and $\mathbf{\Gamma} = \text{diag}(\gamma_i)$ contains the ordered eigenvalues. This is useful to consider since the eigenvalues can be used to bound the joint mutual information captured about individual user symbols, as in (2.117),

$$\log_2\left(1 + \rho\gamma_{\min}\right) \leq \mathcal{I}\left(\mathbf{z}_1, \ldots, \mathbf{z}_L; x_k\right) \leq \log_2\left(1 + \rho\gamma_{\max}\right). \tag{5.29}$$

For $\gamma_{\min} > 0$ a full rank channel is required, meaning the total signal dimension must be greater

than the number of users, $N \geq K/L$. Using Poincare's separation theorem it is straightforward to show that the eigenvalues are upper bounded by the respective eigenvalues of $\mathbf{H}_{\mathrm{G}}$, $\gamma_i \leq \lambda_{\mathrm{G},i}$. Thus for poorly chosen dimension reduction filters the reduced dimension channel eigenvalue spread, $\kappa = \gamma_{\max}/\gamma_{\min}$ may be large. However, the T-CKLT method will tend to act at each stage to increase the smaller eigenvalues, since the $\mathbf{Q}_l$ matrix weights the eigenvectors of $\mathbf{H}_l\mathbf{Q}_l\mathbf{H}_l^{\dagger}$ towards the signal space of the weaker channel eigenvalues. Figure 5.5 shows the CDF of the reduced dimension channel eigenvalue spread using T-CKLT dimension reduction and using random dimension reduction filters. For any signal dimension, the T-CKLT method produces

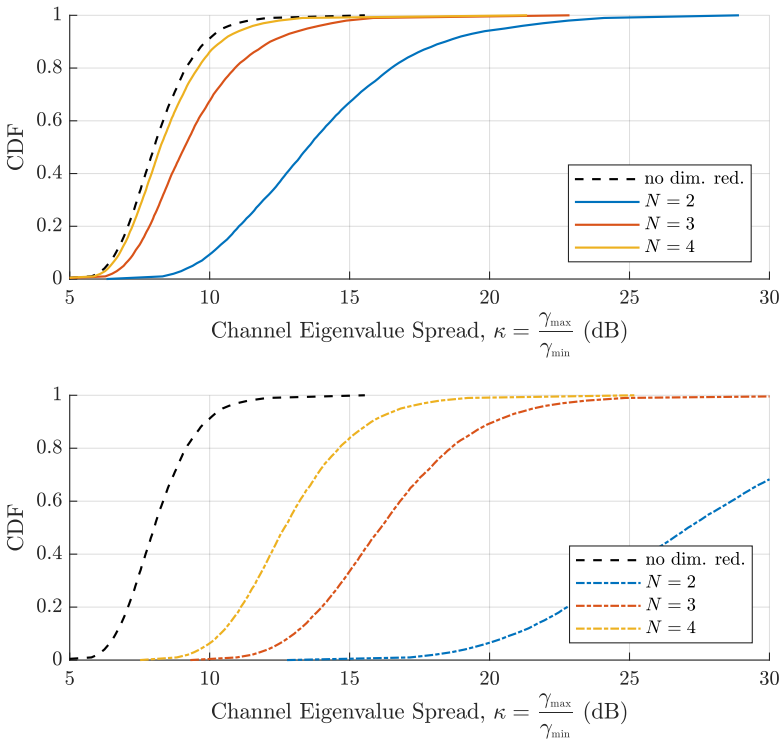

Figure 5.5: Reduced dimension channel eigenvalue spread, $K = 8$, $M = 8$, $L = 4$ ($t = 8$). Top: T-CKLT filters, bottom: random semi-orthogonal filters.

much smaller eigenvalue spreads than random filtering. However, with the minimum signal dimension, $N = K/L$ ($= 2$), optimal filtering will still often produce a poorly conditioned channel, whilst for $N > K/L$ the eigenvalue spread will be close to that of the full dimension channel. This indicates that the total number of signal dimensions needs to be larger than K to achieve a good representation of the received signal.

Benefit of Additional Antennas

When the full dimension signals are used, adding more antennas to each receiver provides array gain and channel hardening. For the reduced dimension case where N is fixed, the benefits of deploying additional antennas are less clear.

With random dimension reduction filtering, it is straightforward to see that deploying additional antennas has no benefit – in a random signal coordinate basis, the received information

will be shared equally (on average) between all signal dimensions. However, the signal does become more concentrated in some *specific* signal basis as the number of antennas is increased[3]. The optimised dimension reduction filter chooses the optimal signal basis and can therefore extract some array gain as M increases. This is illustrated in the example in Figure 5.6 with fixed $N = 3$, where a doubling of the number of antennas provides a 3 dB array gain. This indicates that with a fixed signal dimension, additional receive antennas can be used to reduce the user transmit power. Similarly, a channel hardening effect can also be observed as M increases,

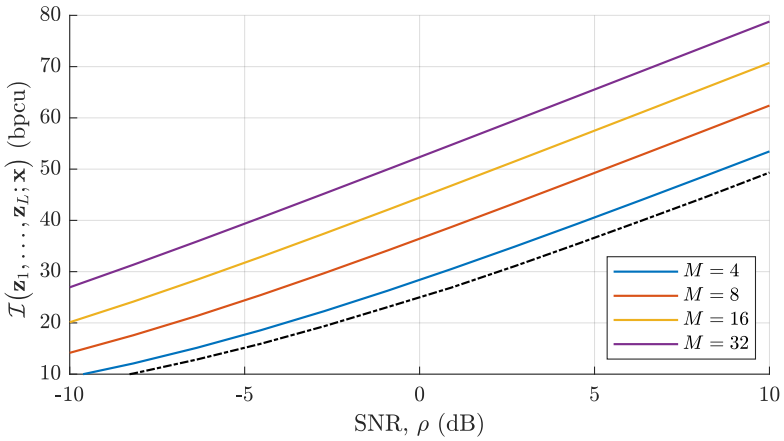

Figure 5.6: Array gain under dimension reduction, $N = 3$, $K = 8$, $L = 4$. Solid line: T-CKLT, dot-dash line: random filtering (all M).

as shown by the eigenvalue distributions in Figure 5.7. This again contrasts with the random filtering case, which does not benefit from channel hardening.

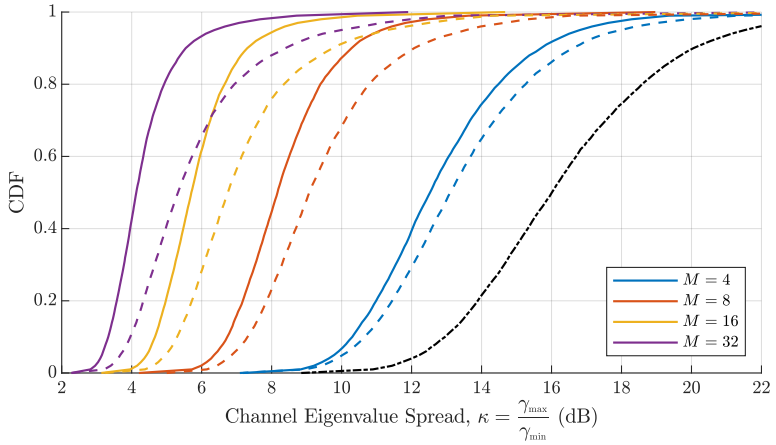

Figure 5.7: Channel hardening under dimension reduction, $M = 8$, $K = 8$, $L = 4$. Solid line: T-CKLT, $N = 4$, dashed line: T-CKLT, $N = 3$, dot-dash line: random filtering, $N = 3$ (all values of M).

[3]This follows from Weyl's inequality [91], since adding an additional antenna increases at least some of the local channel eigenvalues, with none decreasing

$$\mathrm{eig}\big(\mathbf{H}_l^\dagger \mathbf{H}_l\big)_i \leq \mathrm{eig}\big(\mathbf{H}_l^\dagger \mathbf{H}_l + \mathbf{h}_{M+1}\mathbf{h}_{M+1}^\dagger\big)_i.$$

5.3.3 Matched Filter-based Distributed Dimension Reduction

Other distributed dimension reduction strategies aside from the optimal T-CKLT method are also capable of exploiting CSI to capture the benefits of having an excess of receive antennas. As alluded to earlier, for example, antenna selection can be seen as a simple dimension reduction strategy, and is known to be effective in the single receiver setting [71]. Whilst antenna reduction can also be applied in a distributed setting, this section proposes a novel dimension reduction scheme that more directly exploits the characteristics of distributed MIMO networks.

The physical distribution of users and receivers in a distributed MIMO system means that each receiver has strong channels to some users and weak channels to others. If each receiver filters its signal in the 'direction' of a subset of N users, using a set of matched filters, then its reduced dimension signal will capture array gain and diversity for at least those N users. With each receiver focusing on subsets of users to which it has good channels – with the user subsets for all receivers jointly selected at the CP – a good joint signal representation can be expected.

This approach is attractive from a signalling overhead perspective, since only the indices of the selected user vectors are required by the receivers to reconstruct the filters (assuming each receiver already has access to its own CSI), compared to the T-CKLT filters which must be calculated centrally and fed back. This approach has certain similarities to the downlink sparse beamforming [48], with the distinction that here the reduced dimension signals contain uplink signal components from all users, not just the selected users (since the matched filtering does not generally eliminate all inter-user interference[4]).

The channel matrix associated with the N selected user vectors at receiver l is

$$\mathbf{H}_l^{(\mathcal{S}_l)} = \begin{bmatrix} \mathbf{h}_{l,\mathcal{S}_l(1)} & \dots & \mathbf{h}_{l,\mathcal{S}_l(N)} \end{bmatrix}. \tag{5.30}$$

where $\mathcal{S}_l$ is the set of indices of the selected vectors. This reduced channel matrix will generally not have orthonormal columns, and therefore an equivalent semi-orthogonal filter may instead be defined based on the QR decomposition

$$\mathbf{H}_l^{(\mathcal{S}_l)} = \mathbf{A}_l \mathbf{T}_l \tag{5.31}$$

where $\mathbf{A}_l$ has orthonormal columns and $\mathbf{T}_l$ is upper triangular. Using $\mathbf{A}_l$ as a receive filter captures the same information as matched filtering with $\mathbf{H}_l^{(\mathcal{S}_l)}$.

The receive filter $\mathbf{A}_l$ has columns

$$\mathbf{A}_l = \begin{bmatrix} \mathbf{a}_{l,1} & \dots & \mathbf{a}_{l,N} \end{bmatrix} \tag{5.32}$$

which can be calculated sequentially using the Gram-Schmidt procedure

$$\mathbf{a}_{l,i} = \frac{\mathbf{P}_{l,i}\mathbf{h}_{l,\mathcal{S}_l(i)}}{\|\mathbf{P}_{l,i}\mathbf{h}_{l,\mathcal{S}_l(i)}\|} \tag{5.33}$$

[4]The matched filtering here is not being used for detection in the manner described in 2.3.1, but rather to produce a good joint signal representation from which the user symbols can later be jointly detected.

where $\mathbf{P}_{l,i}$ is an orthogonal projection matrix

$$\mathbf{P}_{l,i} = \mathbf{I}_M - \sum_{j<i} \mathbf{a}_{l,j}\mathbf{a}_{l,j}^{\dagger}. \tag{5.34}$$

The joint mutual information can be expressed in terms of these column vectors as

$$\mathcal{I}\big(\mathbf{z}_1, \ldots, \mathbf{z}_L; \mathbf{x}\big) = \log_2 \det\big(\mathbf{I}_K + \rho \sum_{l=1}^{L}\sum_{i=1}^{N} \mathbf{H}_l^{\dagger}\mathbf{a}_{l,i}\mathbf{a}_{l,i}^{\dagger}\mathbf{H}_l\big). \tag{5.35}$$

Finding the optimal sets of user vectors for all receivers is a combinatorial problem with $\binom{K}{N}^L$ combinations, and hence an exhaustive search is prohibitive. Inspired by the antenna selection scheme in [65], a more tractable approach is to instead use a greedy algorithm to select the user matched filtering vectors one at a time, maximising the mutual information at each selection stage.

MF-GS Greedy Selection Algorithm

If after n stages of the greedy algorithm the partially constructed set of MF vectors at receiver l is $\mathcal{S}_l^{(n)}$ and $\mathbf{z}_l^{(n)}$ is the partially constructed reduced dimension signal, with $\mathbf{z}^{(n)} = \{\mathbf{z}_1^{(n)} \ \ldots \ \mathbf{z}_L^{(n)}\}$, then the joint mutual information is

$$\mathcal{I}\big(\mathbf{z}^{(n)}; \mathbf{x}\big) = \log_2 \det\big(\mathbf{I}_K + \rho \sum_{l=1}^{L}\sum_{i=1}^{|\mathcal{S}_l^{(n)}|} \mathbf{H}_l^{\dagger}\mathbf{a}_{l,i}\mathbf{a}_{l,i}^{\dagger}\mathbf{H}_l\big). \tag{5.36}$$

This can be expanded using the conditional mutual information

$$\mathcal{I}\big(\mathbf{z}^{(n)}; \mathbf{x}\big) = \mathcal{I}\big(\mathbf{z}^{(n)}; \mathbf{x}\big|\mathbf{z}^{(n-1)}\big) + \mathcal{I}\big(\mathbf{z}^{(n-1)}; \mathbf{x}\big), \tag{5.37}$$

with the increase in mutual information at stage n

$$\mathcal{I}\big(\mathbf{z}^{(n)}; \mathbf{x}\big|\mathbf{z}^{(n-1)}\big) = \log_2\big(1 + \rho\mathbf{a}_{l,i}^{\dagger}\mathbf{H}_l\mathbf{Q}_{n-1}\mathbf{H}_l^{\dagger}\mathbf{a}_{l,i}\big) \tag{5.38}$$

where $\mathbf{a}_{l,i}$ is the filter vector selected at stage n, and

$$\mathbf{Q}_{n-1} = \big(\mathbf{I}_K + \rho \sum_{l=1}^{L}\sum_{i=1}^{|\mathcal{S}_l^{(n-1)}|} \mathbf{H}_l^{\dagger}\mathbf{a}_{l,i}\mathbf{a}_{l,i}^{\dagger}\mathbf{H}_l\big)^{-1}. \tag{5.39}$$

Substituting (5.33), the joint mutual information is maximised at stage n by choosing the user vector at receiver l that maximises

$$\mathbf{a}_{l,i} = \underset{\mathbf{h}_{l,k}}{\arg}\ \underset{k\notin\mathcal{S}_l^{(n-1)}}{\text{maximize}}\quad \frac{\mathbf{h}_{l,k}^{\dagger}\mathbf{P}_{l,i}\mathbf{H}_l\mathbf{Q}_{n-1}\mathbf{H}_l^{\dagger}\mathbf{P}_{l,i}\mathbf{h}_{l,k}}{\|\mathbf{P}_{l,i}\mathbf{h}_{l,k}\|^2}. \tag{5.40}$$

The $\mathbf{Q}_{n-1}$ matrix can then be updated using a rank-1 update [170]

$$\mathbf{Q}_n = \mathbf{Q}_{n-1} - \frac{\mathbf{Q}_{n-1}\mathbf{H}_l^\dagger \mathbf{a}_{l,i}\mathbf{a}_{l,i}^\dagger \mathbf{H}_l \mathbf{Q}_{n-1}}{1/\rho + \mathbf{a}_{l,i}^\dagger \mathbf{H}_l^\dagger \mathbf{Q}_{n-1}\mathbf{H}_l \mathbf{a}_{l,i}}. \tag{5.41}$$

This greedy selection can be carried out in a round-robin manner, selecting a MF vector for each receiver in turn, as shown in Algorithm 5. This is referred to herein as the matched-filter Gram-Schmidt (MF-GS) filter design algorithm.

Algorithm 5 MF-GS Algorithm

inputs: $\mathbf{H}_l \quad \forall l$

$\mathbf{Q} \leftarrow \mathbf{I}_K$

$\mathbf{P}_l \leftarrow \mathbf{I}_M \quad \forall l$

$\mathcal{S}_l[1:N] \leftarrow 0 \quad \forall l$

for $n = 1 : N$ **do**

  for $l = 1 : L$ **do**

    $k^\star \leftarrow \underset{k}{\arg}\ \underset{k \notin \mathcal{S}_l}{\text{maximize}} \quad \dfrac{\mathbf{h}_{l,k}^\dagger \mathbf{P}_l \mathbf{H}_l \mathbf{Q} \mathbf{H}_l^\dagger \mathbf{P}_l \mathbf{h}_{l,k}}{\mathbf{h}_{l,k}^\dagger \mathbf{P}_l \mathbf{h}_{l,k}}$

    $\mathbf{a} \leftarrow \dfrac{\mathbf{P}_l \mathbf{h}_{l,k^\star}}{\|\mathbf{P}_l \mathbf{h}_{l,k^\star}\|}$

    $\mathcal{S}_l[n] \leftarrow k^\star$

    $\mathbf{Q} \leftarrow \mathbf{Q} - \dfrac{\mathbf{Q}\mathbf{H}_l^\dagger \mathbf{a}\mathbf{a}^\dagger \mathbf{H}_l \mathbf{Q}}{1/\rho + \mathbf{a}^\dagger \mathbf{H}_l^\dagger \mathbf{Q}\mathbf{H}_l \mathbf{a}}$

    $\mathbf{P}_l \leftarrow \mathbf{P}_l - \mathbf{a}\mathbf{a}^\dagger$

  end for

end for

outputs: $\mathcal{S}_l \quad \forall l$

The MF-GS dimension reduction design method requires only matrix multiplications, and the computational complexity is dominated by calculation of $\mathbf{H}_l\mathbf{Q}\mathbf{H}_l^\dagger$. The overall complexity order is $\mathcal{O}\big(K^2LMN\big)$, and fewer overall computations are required compared to the T-CLKT BCA method since no matrix inversions or eigenvalue decompositions are required. Ideas from [65] can be applied to further reduce computational complexity. The method also has the advantage of producing the same first n outputs for any N, and can therefore be used to find dimension reduction filters for various values of N with minimal additional computation.

Figure 5.8 compares the performance of the MF-GS method to the T-CKLT method and a simple greedy antenna selection algorithm adapted from [65] (not shown here). The MF-GS method significantly outperforms antenna selection, and comes close to the performance of the T-CKLT method. Comparing to Figure 5.5, all methods significantly improve performance compared to using a reduced number of antennas or a random dimension reduction filter.

Whilst the MF-GS filters produce less accurate signal representations than the T-CKLT filters, they have the benefit of reducing signalling overheads since only the N indices of the selected user vectors need to be fed back to each receiver for local filter reconstruction, as opposed to the full $\mathbf{A}_l$ matrix. Assuming local CSI is initially obtained at the receivers using uplink pi-

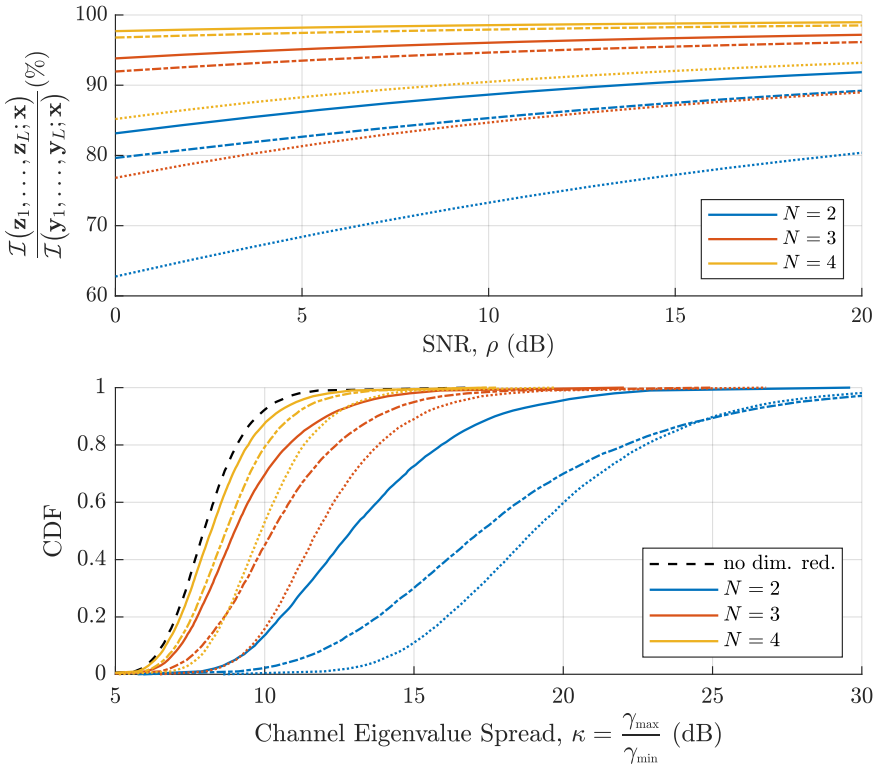

Figure 5.8: Performance of different dimension reduction schemes, $K = 8$, $L = 4$. Solid line: T-CKLT, dot-dash line: MF-GS, dotted line: antenna selection.

lots, the signalling overheads under MF-GS distributed dimension reduction are dominated by the transfer of CSI to the CP. When the channel changes regularly due to fading, these CSI overheads can themselves be reduced by adapting the MF-GS method.

Extension for Block Fading Channels

For block fading channels, the channel matrix realisations are randomly drawn according to the channel statistics, and then assumed to be constant for all transmissions within a coherence block. The ergodic, or average, mutual information (over multiple coherence blocks) is then defined

$$\mathbb{E}\big[\mathcal{I}(\mathbf{z}_1,\ldots,\mathbf{z}_L;\mathbf{x})\big] = \mathbb{E}\big[\log_2\det\big(\mathbf{I}_K + \rho\sum_{l=1}^{L}\mathbf{H}_l^\dagger\mathbf{A}_l\mathbf{A}_l^\dagger\mathbf{H}_l\big)\big]. \tag{5.42}$$

For good performance, the dimension reduction filters, $\mathbf{A}_l$, should vary with the random channel realisation, $\mathbf{H}_l$. With both the T-CKLT and MF-GS methods, this means filter design must be performed at the CP at each channel coherence interval, requiring full CSI be transferred to the CP (KM coefficients per receiver). However, once dimension reduction has been applied at the receivers, any processing of the reduced dimension signals only requires knowledge of the reduced dimension channel matrices $\mathbf{G}_l$ (KN coefficients per receiver).

In the MF-GS scheme the filters are parameterised by the set of indices of the selected MF vectors, $\mathcal{S}_l$. The set of users that is selected can be expected to be strongly influenced by

the path-loss, or slow fading characteristics, of the user channels. Since the path-loss changes slowly relative to the fast fading channel realisation, rather than choosing the optimal $\mathcal{S}_l$ for each individual channel realisation, $\mathcal{S}_l$ can instead be fixed for a group of coherence blocks over which the slow fading characteristics remain approximately constant. The receive filters may then be updated locally at each receiver at each coherence interval using local knowledge of $\mathbf{H}_l$, and only the reduced channel matrices $\mathbf{G}_l$ need to then be transferred to the CPU for any processing, rather than full CSI.

Adapting the original MF-GS algorithm, the $\mathcal{S}_l$ can be found by greedily selecting the indices that maximise the average joint mutual information, i.e. at stage n

$$k^\star = \arg\max_k \quad \mathbb{E}\Big[\log_2\Big(1+\frac{\rho \mathbf{h}_{l,k}^\dagger \mathbf{P}_{l,i}\mathbf{H}_l\mathbf{Q}_{n-1}\mathbf{H}_l^\dagger\mathbf{P}_{l,i}\mathbf{h}_{l,k}}{\mathbf{h}_{l,k}^\dagger\mathbf{P}_{l,i}\mathbf{h}_{l,k}}\Big)\Big] \tag{5.43}$$

where $\mathbf{h}_{l,k}$, $\mathbf{P}_{l,i}$ and $\mathbf{Q}_{n-1}$ are all random quantities that vary with the channel realisations. The $\mathcal{S}_l$ can be chosen for arbitrary channel distributions using a sample of N_s channel realisations, $\mathbf{H}_l^{(i)}$, taken from different coherence blocks over which the channel slow fading statistics stay approximately constant. For example, the channel realisation samples could be obtained using pilots on sufficiently separated subcarriers in an OFDM symbol. The fading MF-GS algorithm (F-MF-GS) is shown in Algorithm 6.

Algorithm 6 F-MF-GS Receive Filter Design for Fading Channels

inputs: $\mathbf{H}_l^{(i)} \quad \forall l,i$
$\mathbf{Q}^{(i)} \leftarrow \mathbf{I}_K \quad \forall i$
$\mathbf{P}_l^{(i)} \leftarrow \mathbf{I}_M \quad \forall l,i$
$\mathcal{S}_l[1:N] \leftarrow 0 \quad \forall l$
for $n = 1:N$ **do**
  for $l = 1:L$ **do**
    $\gamma_k \leftarrow \frac{1}{N_s}\sum_i \log_2\left(1+\frac{\rho \mathbf{h}_{l,k}^{(i)\dagger}\mathbf{P}_l^{(i)}\mathbf{H}_l^{(i)}\mathbf{Q}^{(i)}\mathbf{H}_l^{(i)\dagger}\mathbf{P}_l^{(i)}\mathbf{h}_{l,k}^{(i)}}{\mathbf{h}_{l,k}^{(i)\dagger}\mathbf{P}_l^{(i)}\mathbf{h}_{l,k}^{(i)}}\right) \quad \forall k \notin \mathcal{S}_l$
    $k^\star \leftarrow \arg\max_k \quad \gamma_k$
    $\mathcal{S}_l[n] \leftarrow k^\star$
    $\mathbf{a}^{(i)} \leftarrow \frac{\mathbf{P}_l^{(i)}\mathbf{h}_{l,k^\star}^{(i)}}{\|\mathbf{P}_l^{(i)}\mathbf{h}_{l,k^\star}^{(i)}\|} \quad \forall i$
    $\mathbf{Q}^{(i)} \leftarrow \mathbf{Q}^{(i)} - \frac{\mathbf{Q}^{(i)}\mathbf{H}_l^{(i)\dagger}\mathbf{a}^{(i)}\mathbf{a}^{(i)\dagger}\mathbf{H}_l^{(i)}\mathbf{Q}^{(i)}}{1/\rho + \mathbf{a}^{(i)\dagger}\mathbf{H}_l^{(i)\dagger}\mathbf{Q}^{(i)}\mathbf{H}_l^{(i)}\mathbf{a}^{(i)}} \quad \forall i$
    $\mathbf{P}_l^{(i)} \leftarrow \mathbf{P}_l^{(i)} - \mathbf{a}^{(i)}\mathbf{a}^{(i)\dagger} \quad \forall i$
  end for
end for
outputs: $S_l \quad \forall l$

The performance degradation compared to the full MF-GS method is small, as shown in Figure 5.9.

The adapted MF-GS method for fading channels reduces the signalling overheads even com-

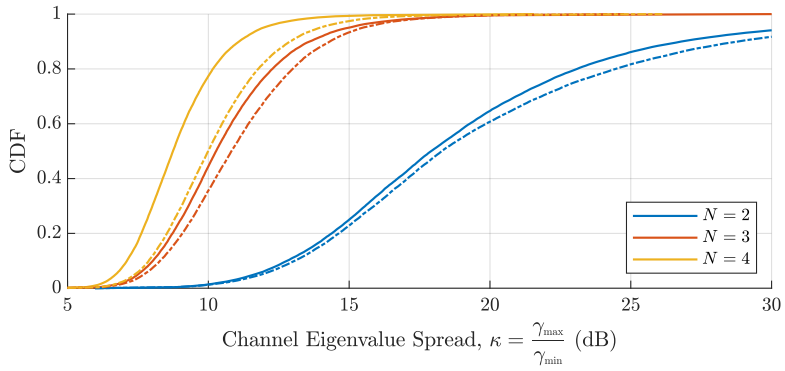

Figure 5.9: Channel eigenvalue spread under dimension reduction, $K = 8$, $M = 8$, $L = 4$. Solid line: MF-GS, dashed line: F-MF-GS, $N_s = 4$.

pared to the case where no dimension reduction is used since only the reduced dimension channels, $\mathbf{G}_l$, need to be transferred over fronthaul, rather than full dimension channels, $\mathbf{H}_l$. Figure 5.10 illustrates the saving for the case where $N_s = 4$ channel matrix measurements are taken from 4 different coherence blocks within the operating frequency bandwidth and used to choose the $\mathcal{S}_l$. The selected indices are held constant over $N_c = 32$ coherence blocks, meaning that for the 28 remaining coherence blocks dimension reduction can be performed at the receivers, with only the reduced dimension channels transferred over fronthaul.

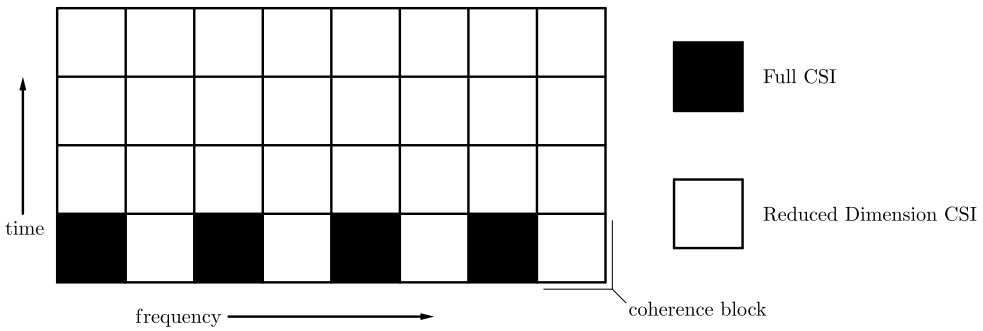

Figure 5.10: Illustration of reduced CSI signalling overheads facilitated by F-MF-GS dimension reduction scheme.

Assuming each element in $\mathbf{H}_l$, $\mathbf{A}_l$ and $\mathbf{G}_l$ is represented using n_b bits, Table 5.3.3 shows the signalling overheads per coherence block associated with the different schemes. The proportion of fronthaul capacity that must be devoted to signalling overheads then depends on the channel coherence block size – channels with high mobility and rich multipath content having higher overheads due to the more frequent CSI updates that are required.

	Average Signalling Overheads (bits per coherence block)			
Filter Design Scheme	RRH to CP		CP to RRH	
No Dimension Reduction	$\mathbf{H}_l$	MKn_b	-	-
T-CKLT	$\mathbf{H}_l$	MKn_b	$\mathbf{A}_l$	MNn_b
MF-GS	$\mathbf{H}_l$	MKn_b	$\mathcal{S}_l$	$N\log_2 K$
F-MF-GS	$\mathbf{H}_l/\mathbf{G}_l$	$\frac{1}{N_c}\big(MN_s+N(N_c-N_s)\big)Kn_b$	$\mathcal{S}_l$	$\frac{1}{N_c}N\log_2 K$

The overheads are illustrated in Figure 5.11 for a system with $M = 8, K = 8$, coherence block length of 100 symbols and $n_b = 20$ bits.

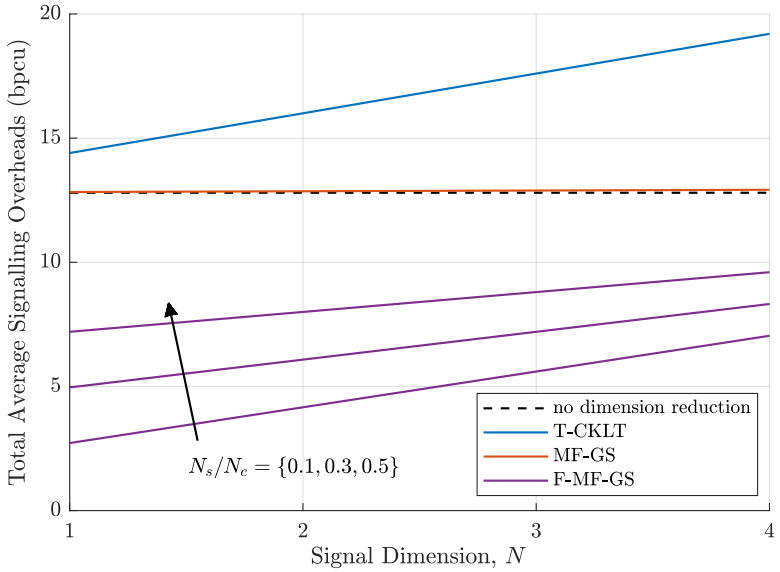

Figure 5.11: Average signalling overheads associated with different dimension reduction schemes, channel coherence block size 100 symbols, $K = 8$, $M = 8$, $n_b = 20$ bits.

5.3.4 Dimension Reduction with Imperfect CSI

Practical systems must estimate CSI using transmitted pilot signals, and hence suffer from imperfect CSI. Under the MMSE channel estimation model, it is straightforward to extend the dimension reduction filter design methods to account for this. As in Section 4.3.5, a whitening transform is first applied,

$$\begin{aligned}\check{\mathbf{y}}_l &= \mathbf{\Omega}_l^{-1/2}\mathbf{y}_l \\ &= \check{\mathbf{H}}_l\mathbf{x} + \check{\boldsymbol{\omega}}_l,\end{aligned} \tag{5.44}$$

before dimension reduction

$$\mathbf{z}_l = \mathbf{A}_l^{\dagger}\check{\mathbf{y}}_l. \tag{5.45}$$

Treating the signal through the CSI errors as noise, the average joint mutual information can be lower bounded

$$\mathbb{E}_{\mathbf{E}}\big[\mathcal{I}\big(\mathbf{z}_1,\ldots,\mathbf{z}_L;\mathbf{x}\big)\big] \geq \mathbb{E}_{\mathbf{E}}\Big[\log_2\det\Big(\mathbf{I}_K + \rho\sum_{l=1}^{L}\check{\mathbf{H}}_l^{\dagger}\mathbf{A}_l\mathbf{A}_l^{\dagger}\check{\mathbf{H}}_l\Big)\Big]. \tag{5.46}$$

This has the same form as (5.11), and is maximised by applying the T-CKLT BCA algorithm to the whitened channel estimates $\check{\mathbf{H}}_l = \boldsymbol{\Omega}_l^{-1/2}\hat{\mathbf{H}}_l$. For spatially correlated user fading channels, this whitening transformation, $\boldsymbol{\Omega}_l^{-1/2}$, can be seen as a weighting that causes the receive filters to favour signal subspaces which (on average) contain lower channel estimation error.

The MF-GS algorithm may similarly be applied using the modified channel vectors, $\check{\mathbf{h}}_{l,k} = \boldsymbol{\Omega}_l^{-1/2}\hat{\mathbf{h}}_{l,k}$. The CP only requires knowledge of the transformed channel, $\check{\mathbf{H}}_l$, and hence no additional signalling overheads are required compared to the perfect CSI case.

As before, a reduced dimension estimated channel can be defined

$$\hat{\mathbf{G}}_l = \mathbf{A}_l^{\dagger}\boldsymbol{\Omega}_l^{-1/2}\hat{\mathbf{H}}_l. \tag{5.47}$$

The performance of the reduced dimension channel with imperfect CSI is interference limited, and therefore the bound stops growing at high SNR due to the CSI errors,

$$\lim_{\rho\to\infty} \log_2 \det\Big(\mathbf{I}_K + \rho\sum_{l=1}^{L}\hat{\mathbf{G}}_l^{\dagger}\hat{\mathbf{G}}_l\Big) = \log_2 \det\Big(\mathbf{I}_K + \sum_{l=1}^{L}\hat{\mathbf{H}}_l^{\dagger}\Big(\sum_{k=1}^{K} p_k\mathbf{C}_{l,k}\Big)^{-1/2}\mathbf{A}_l\mathbf{A}_l^{\dagger}\Big(\sum_{k=1}^{K} p_k\mathbf{C}_{l,k}\Big)^{-1/2}\hat{\mathbf{H}}_l\Big). \tag{5.48}$$

This constrasts with the perfect CSI case, where increasing the SNR increases the joint mutual information. As the quality of the channel estimates increases (the $\mathbf{C}_{l,k}$ decrease) the joint mutual information lower bound increases as shown in Figure 5.12.

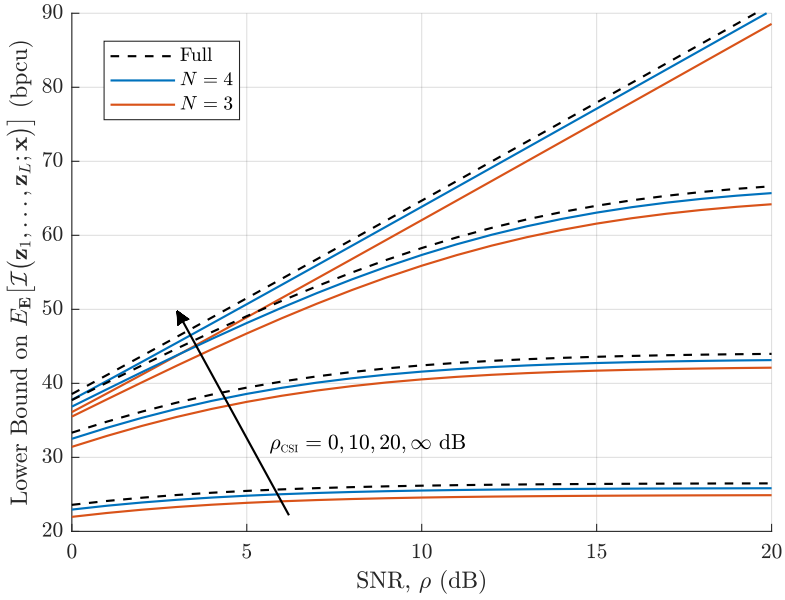

Figure 5.12: Joint mutual information lower bound under T-CKLT dimension reduction, $K = 8, M = 8, L = 4$.

5.4 Reduced Dimension MIMO Uplink with Lossy Compression

So far it has been shown that lossy distributed dimension reduction can exploit the joint sparsity of the received signals in a distributed MIMO network with an excess of receive antennas to produce reduced dimension signal representations that accurately preserve the salient features. However, before these signals can be transferred over finite-capacity fronthaul, lossy signal compression must be applied to encode them using a finite number of bits.

In contrast to the previous chapter, where the use of transform coding with a jointly optimised rate allocation was studied, this section considers a simple compression scheme where the reduced dimension signals are directly compressed using N equal-resolution scalar compressors/quantizers. Since this lossy compression stage does not adapt to the signal statistics, all improvements in fronthaul utilisation come from the distributed dimension reduction stage.

First, the use of Gaussian scalar compression is considered. A block coordinate ascent algorithm for finding the sum capacity maximising dimension reduction filters is derived, before the sum capacity scaling in the high SNR regime is analysed. Comparing dimension reduction to the optimal point-to-point compression outlined in Chapter 4, it is argued that at high SNR dimension reduction is a quasi-optimal point-to-point compression strategy for maximising fronthaul efficiency.

A practical dimension reduction compression strategy for distributed MIMO networks is then proposed, which uses the dimension reduction filters from Section 5.3 in conjunction with simple fixed-rate scalar quantization.

5.4.1 Sum Capacity under Gaussian Scalar Compression

Applying Gaussian scalar compression directly to the reduced dimension signals using a simple uniform rate allocation, $r_i = \mathcal{R}/N$, produces compressed signals

$$\tilde{\mathbf{z}}_l = \mathbf{G}_l\mathbf{x} + \boldsymbol{\eta} + \boldsymbol{\delta}_l \tag{5.49}$$

with quantization noise $\boldsymbol{\delta}_l \sim \mathcal{CN}\big(0, \boldsymbol{\Phi}_l\big)$, where

$$\boldsymbol{\Phi}_l = \big(\rho\mathbf{D}_l + \mathbf{I}_N\big)\frac{1}{2^{\mathcal{R}/N} - 1} \tag{5.50}$$

with $\mathbf{D}_l = \operatorname{diag}\left(\|\mathbf{g}_{l,i}\|^2\right)_{i=1}^{N}$ and $\mathbf{g}_{l,i} = \mathbf{H}_l^{\dagger}\mathbf{f}_{l,i}$.

Comparing with Section 4.2.2, it is clear that – like the transform coding scheme investigation in Section 4.4 – the reduced dimension compression scheme is a specialised case of Gaussian vector point-to-point compression. However, here efficient compression is achieved by optimising the dimension reduction filters applied at the receivers, in contrast to the transform coding approach where a fixed transform is used in conjunction with optimised rate allocation.

Sum Capacity Maximising Dimension Reduction

The maximum fronthaul efficiency is achieved by maximising the sum capacity under MMSE-SIC detection with respect to the transforms,

$$\underset{\mathbf{A}_1,\ldots,\mathbf{A}_L}{\text{maximise}} \quad \log_2 \det\big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \mathbf{H}_l^\dagger \mathbf{A}_l \big(\mathbf{I}_N + \mathbf{\Phi}_l\big)^{-1} \mathbf{A}_l^\dagger \mathbf{H}_l\big) \tag{5.51}$$

$$\text{subject to} \quad \mathbf{A}_l^\dagger \mathbf{A}_l = \mathbf{I}_N \quad \forall l. \tag{5.52}$$

This is a non-convex problem and hence it is not possible to find a global maximum, but a block coordinate ascent may again be used to find a stationary point. Here the filter columns, $\mathbf{a}_{l,i}$, are updated in turn using the mutual information expansion

$$\mathcal{C}_{\text{SUM}} = \mathcal{I}\big(\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_L; \mathbf{x}\big) \tag{5.53}$$

$$= \mathcal{I}\big(\tilde{z}_{l,i}; \mathbf{x} \big| \tilde{\mathbf{z}}_{l,i}^{\mathsf{c}}\big) + \mathcal{I}\big(\tilde{\mathbf{z}}_{l,i}^{\mathsf{c}}; \mathbf{x}\big) \tag{5.54}$$

where $\tilde{z}_{l,i}$ is component i at receiver l and $\tilde{\mathbf{z}}_{l,i}^{\mathsf{c}}$ is the set of all other $NL-1$ components. Only the first term depends on $\mathbf{a}_{l,i}$, and can be shown to be

$$\mathcal{I}\big(\tilde{z}_{l,i}; \mathbf{x} \big| \tilde{\mathbf{z}}_{l,i}^{\mathsf{c}}\big) = \log_2 \left(1 + \big(2^{\mathcal{R}/N} - 1\big) \frac{\rho \mathbf{f}_{l,i}^\dagger \mathbf{H}_l \mathbf{T}_{l,i} \mathbf{H}_l^\dagger \mathbf{f}_{l,i}}{\mathbf{f}_{l,i}^\dagger \big(\rho \mathbf{H}_l \mathbf{H}_l^\dagger + 2^{\mathcal{R}/N} \mathbf{I}_M\big) \mathbf{f}_{l,i}}\right) \tag{5.55}$$

where

$$\mathbf{T}_{l,i} = \Big(\mathbf{I}_K + \rho \sum_{j \neq l} \mathbf{G}_j^\dagger \big(\mathbf{I}_N + \mathbf{\Phi}_j\big)^{-1} \mathbf{G}_j + \rho \sum_{u \neq i} \frac{\mathbf{g}_{l,u} \mathbf{g}_{l,u}^\dagger}{\phi_{l,i} + 1}.\Big)^{-1} \tag{5.56}$$

The optimal $\mathbf{a}_{l,i}$ is then the solution to

$$\begin{aligned} \underset{\mathbf{a}_{l,i}}{\text{maximise}} \quad & \frac{\rho \mathbf{a}_{l,i}^\dagger \mathbf{H}_l \mathbf{T}_{l,i} \mathbf{H}_l^\dagger \mathbf{a}_{l,i}}{\mathbf{a}_{l,i}^\dagger \big(\rho \mathbf{H}_l \mathbf{H}_l^\dagger + 2^{\mathcal{R}/N} \mathbf{I}_M\big) \mathbf{a}_{l,i}} \\ \text{subject to} \quad & \mathbf{a}_{l,i}^\dagger \mathbf{a}_{l,j} = 0, \quad \forall j \neq i. \end{aligned} \tag{5.57}$$

The constraint is introduced to ensure orthogonality of the columns of $\mathbf{A}_l$, and can be expressed using a nullspace constraint

$$\mathbf{a}_{l,i} = \big(\mathbf{I}_M - \sum_{j \neq i} \mathbf{a}_{l,j} \mathbf{a}_{l,j}^\dagger\big) \mathbf{a}_{l,i} \tag{5.58}$$

$$= \mathbf{P}_{l,i} \mathbf{a}_{l,i}. \tag{5.59}$$

Substituting, the constrained problem is converted into an unconstrained generalised Rayleigh quotient problem

$$\underset{\mathbf{a}_{l,i}}{\text{maximise}} \quad \frac{\rho \mathbf{a}_{l,i}^\dagger \mathbf{P}_{l,i} \mathbf{H}_l \mathbf{T}_{l,i} \mathbf{H}_l^\dagger \mathbf{P}_{l,i} \mathbf{a}_{l,i}}{\mathbf{a}_{l,i}^\dagger \mathbf{P}_{l,i} \big(\rho \mathbf{H}_l \mathbf{H}_l^\dagger + 2^{\mathcal{R}/N} \mathbf{I}_M\big) \mathbf{P}_{l,i} \mathbf{a}_{l,i}}, \tag{5.60}$$

that can be solved as a generalised eigenproblem [139],

$$\mathbf{P}_{l,i}\mathbf{H}_l\mathbf{T}_{l,i}\mathbf{H}_l^\dagger\mathbf{P}_{l,i} = \zeta\mathbf{P}_{l,i}\big(\mathbf{H}_l\mathbf{H}_l^\dagger + 2^{\mathcal{R}/N}/\rho\mathbf{I}_M\big)\mathbf{P}_{l,i}, \tag{5.61}$$

where ζ is the largest generalised eigenvalue. Using the BCA procedure a stationary point can be found by updating each basis vector in turn, repeating cyclically. Since at each update the Rayleigh quotient has a unique maxima, the Rayleigh quotient block coordinate ascent algorithm (RQ-BCA) converges monotonically towards a stationary point [162].

Figure 5.13 compares the sum capacity achieved by the RQ-BCA dimension reduction filters for different values of N to that achieved by the SCA-P2P method adapted from [248], for a single channel realisation. Remarkably, the envelope of the dimension reduction capacity curves closely matches the capacity curve of the SCA-P2P compression scheme, and at low fronthaul capacities actually outperforms it. As the dimension reduction scheme is a specific case of point-to-point vector compression this is surprising, but is nonetheless entirely feasible – the SCA-P2P scheme finds a stationary point to the point-to-point compression sum capacity maximisation problem, rather than the global maximum.

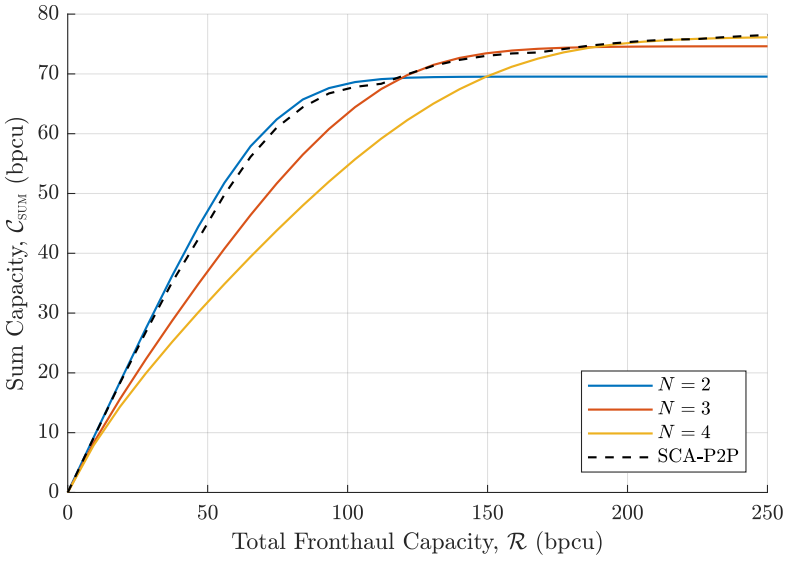

Figure 5.13: Sum capacity under RQ-BCA dimension reduction filters for different signal dimensions, $\rho = 15$ dB, $K = 8, M = 8, L = 4$ (single channel realisation).

This result suggests that *dimension reduction-based compression can be a quasi-optimal strategy for efficiently utilising the fronthaul network in a distributed MIMO network with an excess of receiver antennas.* This observation is consistent with the findings of the previous chapter, where a sparse rate allocation was found to be produced when the rate allocations at different receivers are jointly optimised. Further evidence of this is now provided by analysing the performance of dimension reduction compression at high SNR.

Sum Capacity Scaling at High SNR

Some useful insights can be gained about the performance of dimension reduction-based signal compression scheme by considering the capacity scaling in the high SNR limit – i.e. for the scenario where there is no receiver noise, and capacity is limited only by quantization noise. In this region, the capacity is given by

$$\mathcal{C}_{\text{SUM}}^{\text{SNR}} = \lim_{\rho\to\infty} \mathcal{C}_{\text{SUM}} \tag{5.62}$$

$$= \log_2 \det\left(\mathbf{I}_K + \left(2^{\mathcal{R}/N} - 1\right)\mathbf{\Pi}\right), \tag{5.63}$$

where

$$\mathbf{\Pi} = \sum_{l=1}^{L} \mathbf{G}_l^\dagger \mathbf{D}_l^{-1} \mathbf{G}_l = \sum_{l=1}^{L}\sum_{i=1}^{N} \frac{\mathbf{g}_{l,i}\mathbf{g}_{l,i}^\dagger}{\|\mathbf{g}_{l,i}\|^2}. \tag{5.64}$$

Assuming $\mathbf{\Pi}$ is full rank, the high SNR limit can be simply lower bounded,

$$\mathcal{C}_{\text{SUM}}^{\text{SNR}} > \log_2 \det\left(\left(2^{\mathcal{R}/N} - 1\right)\mathbf{\Pi}\right) \tag{5.65}$$

$$= K \log_2\left(2^{\mathcal{R}/N} - 1\right) + \log_2 \det\left(\mathbf{\Pi}\right), \tag{5.66}$$

whilst using the bound in (4.34) it can be upper bounded

$$\mathcal{C}_{\text{SUM}}^{\text{SNR}} < \log_2 \det\left(\mathbf{I}_K + 2^{\mathcal{R}/N}\mathbf{\Pi}\right) \tag{5.67}$$

$$\leq \frac{\mathcal{R}K}{N} + \log_2 \det\left(\mathbf{\Pi}\right) + 2^{-\mathcal{R}/N}\log_2(e)\,\text{Tr}\left(\mathbf{\Pi}^{-1}\right). \tag{5.68}$$

These two bounds rapidly converge as $\mathcal{R}$ is increased, and for practical $\mathcal{R}/N \gg 1$ the high SNR sum capacity is well approximated by

$$\mathcal{C}_{\text{SUM}}^{\text{SNR}} \approx \frac{\mathcal{R}K}{N} + \log_2 \det\left(\mathbf{\Pi}\right), \tag{5.69}$$

as shown in Figure 5.14 for a network with $K = 8, M = 8, L = 4$.

The sum capacity therefore increases approximately linearly with the per-RRH fronthaul capacity,

$$\frac{\partial \mathcal{C}_{\text{SUM}}^{\text{SNR}}}{\partial \mathcal{R}} \approx \frac{K}{N}, \tag{5.70}$$

scaling most quickly when a small signal dimension is used. For example, with the minimum signal dimension, $N = K/L$, adding 1 bpcu of total fronthaul capacity (shared between L fronthaul connections) increases the sum capacity by 1 bpcu, and performance stays within a gap of the cut-set bound[5]

$$\mathcal{C}_{\text{SUM}}^{\text{SNR}} \approx \mathcal{R}L + \epsilon\,. \tag{5.71}$$

Clearly, no practical distributed MIMO system operates at infinite SNR; however, the approximation is reasonable whenever the quantization noise is much greater than the receiver

[5]It can be shown using Hadamard's inequality that $\epsilon = \log_2 \det\left(\mathbf{\Pi}\right) \leq 0$ when $N = K/L$. For $N > K/L$, $\log_2 \det\left(\mathbf{\Pi}\right)$ may be a positive number.

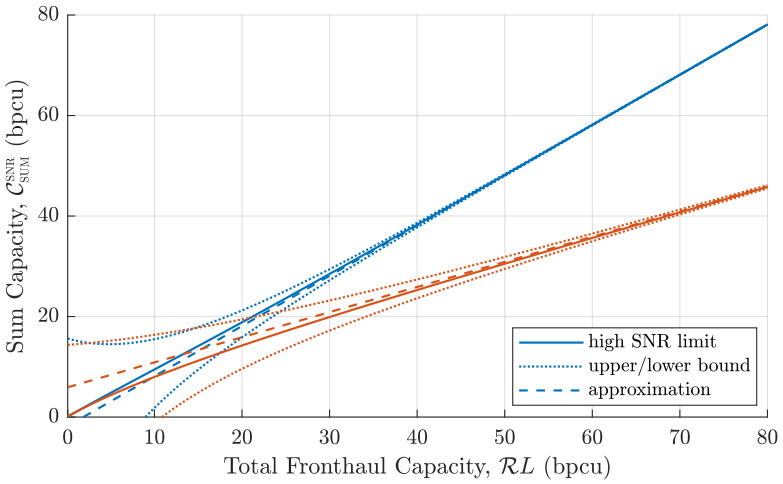

Figure 5.14: Asymptotic high SNR sum capacity scaling under T-CKLT dimension reduction, $K = 8, M = 8, L = 4$. Blue lines: $N = 2$, orange lines: $N = 4$.

noise, $\mathbf{\Phi}_l \gg \mathbf{I}_N$ and is therefore a good approximation at practical high SNRs when operating in the fronthaul-limited region, as illustrated in Figure 5.15 for $\rho = 15$ dB. The assumption of moderately high SNRs is reasonable for dense MIMO C-RAN deployments [234].

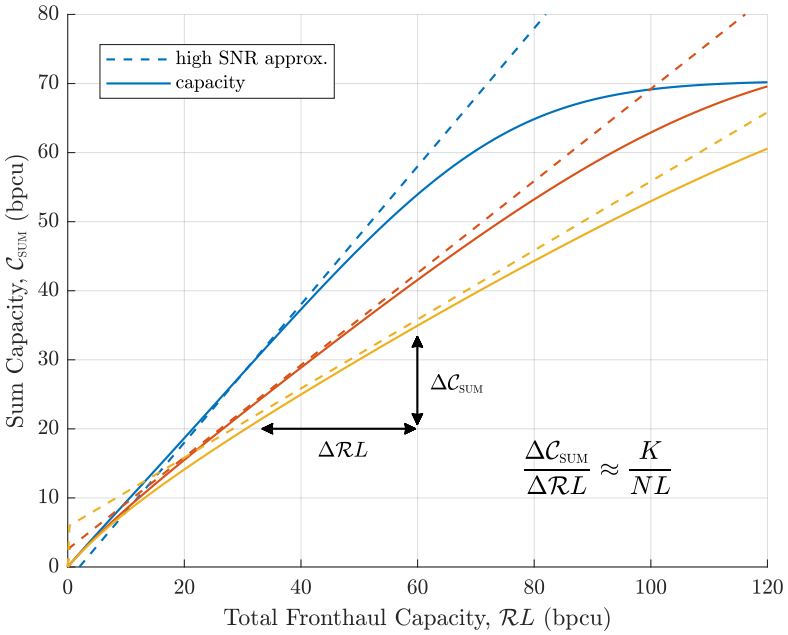

Figure 5.15: Sum capacity scaling under T-CKLT dimension reduction, $\rho = 15$ dB,$K = 8, M = 8, L = 4$. Blue lines: $N = 2$, orange lines: $N = 3$, yellow lines: $N = 4$.

At finite SNR, the quantization noise decreases as the fronthaul capacity increases until

performance is limited instead by receiver noise. When $\mathbf{\Phi}_l \ll \mathbf{I}_N$,

$$\mathcal{C}_{\text{SUM}} \approx \log_2 \det \left(\mathbf{I}_K + \rho \sum_{l=1}^{L} \mathbf{G}_l^\dagger \mathbf{G}_l \right) \tag{5.72}$$

$$= \log_2 \det \left(\mathbf{I}_K + \rho \sum_{l=1}^{L} \mathbf{H}_l^\dagger \mathbf{H}_l \right) - \mathcal{L}, \tag{5.73}$$

and the sum capacity is limited by the information lost from applying dimension reduction. Thus under dimension reduction signal compression two operating regimes can be identified:

- In the quantization noise-limited region the sum capacity increases approximately linearly with the fronthaul capacity. For a given fronthaul capacity the sum capacity (and fronthaul efficiency) is greater when dimension reduction is applied more aggressively – small N – since this results in lower quantization noise.
- In the receiver noise-limited region the sum capacity is limited by the lossy dimension reduction. Here the use of a larger signal dimension can be beneficial since it preserves more of the information from the original signal.

In practice there is a gradual transition between the two regimes, and for a given fronthaul capacity the maximum achievable sum capacity can be found by comparing sum capacities for different values of N. This yields the performance curves in Figure 5.13.

The combination of dimension reduction filtering and scalar compression can achieve high fronthaul efficiency, as shown in Figure 5.16 for two different network configurations. The RQ-BCA, T-CKLT & MF-GS filters all outperform the SCA-RA transform coding scheme at many fronthaul capacities. Whilst only the RQ-BCA dimension reduction filters explicitly account for the effects of quantization noise, the T-CKLT dimension reduction filters achieve effectively the same performance. The use of simple dimension reduction schemes based on antenna selection or random filtering give worse performance, but still provide a capacity gain over UQN compression by using a reduced signal dimension to reduce quantization noise.

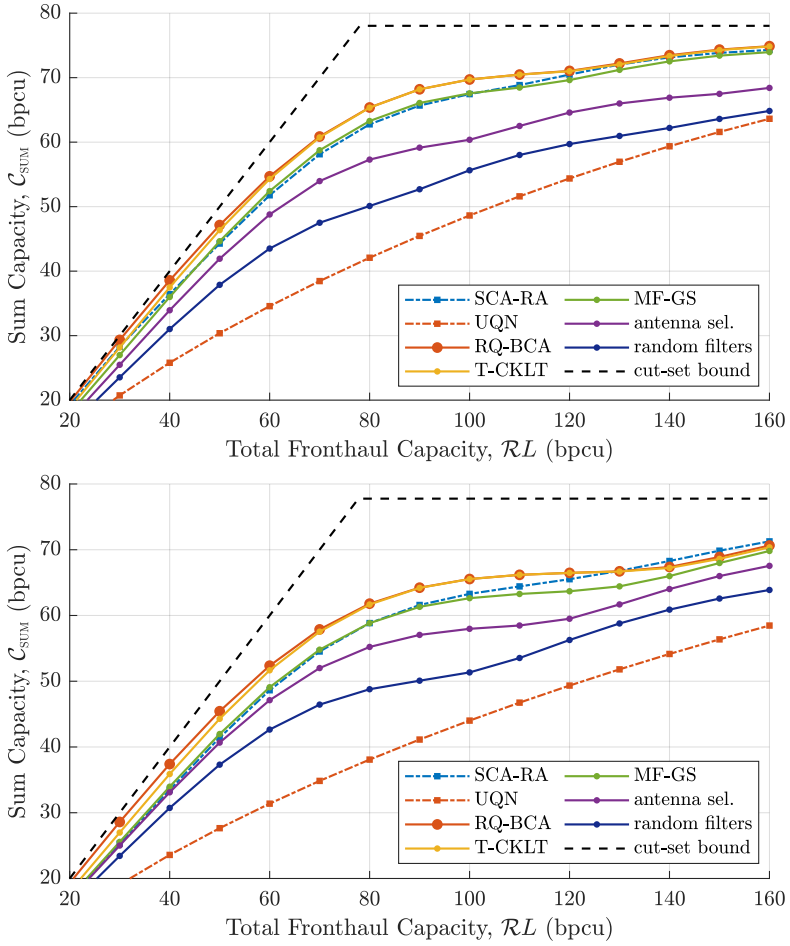

Figure 5.16: Sum capacity performance under different dimension reduction schemes, $\rho = 15$ dB. Top: $K = 8, M = 8, L = 4$, bottom: $K = 8, M = 4, L = 8$.

5.4.2 Practical Reduced Dimension Compression using Fixed-Rate Quantizers

Whilst the study of system sum capacity provides some important insights, in cellular distributed MIMO systems the individual user capacities that can be provided are often of more significance (the sum capacity may be shared unevenly between users). Whilst sum capacity maximising dimension reduction filters can be found as in Section 5.4.1, this approach cannot easily be modified to account for specific per-user capacity criteria. Here, it is elected to instead use filters designed under the joint mutual information criteria in Section 5.3, since these have been shown to provide good reduced dimension signal representations. Furthermore, in practical systems the use of fixed-rate scalar quantization – rather than entropy coded quantization – is attractive for its simplicity. The user capacities achieved by dimension reduction under fixed-rate scalar quantization and linear MMSE detection are therefore now considered.

In the proposed scheme, dimension reduction is applied to the received signals,

$$\mathbf{z}_l = \mathbf{A}_l^{\dagger}\mathbf{y}_l, \tag{5.74}$$

before the reduced dimension signals are quantized using N pairs of equal resolution fixed-rate scalar quantizers, each allocated b quantization bits,

$$\tilde{z}_{l,i} = Q\big(\Re(z_{l,i})\big) + jQ\big(\Im(z_{l,i})\big). \tag{5.75}$$

Following the discussion in Section 4.2.2, using Gaussian scalar Lloyd-Max quantizers, the quantized signals can be modelled using an additive quantization noise model,

$$\tilde{\mathbf{z}}_l = \mathbf{z}_l + \boldsymbol{\delta}_l, \tag{5.76}$$

where $\boldsymbol{\delta}_l$ has unknown distribution but known covariance

$$\mathbb{E}\big[\boldsymbol{\delta}_l \boldsymbol{\delta}_l^\dagger\big] = \boldsymbol{\Phi}_l \approx \frac{\pi\sqrt{3}}{2}\big(\rho\mathbf{D}_l + \mathbf{I}_N\big)2^{-b}. \tag{5.77}$$

This approximation is tight at higher quantization resolutions ($b \geq 8$), and is therefore valid for the case of interest where the $\mathcal{R}$ fronthaul capacity bits are shared between a small number of quantizers[6]. Comparing to (5.50), the use of fixed-rate quantizers (rather than entropy coded) increases the quantization noise by a factor of approximately $\frac{\pi\sqrt{3}}{2}$. Achieving the same quantization noise level therefore requires an extra $1.4N$ bits of fronthaul capacity for each receiver – when N is small, simpler fixed-rate quantization is competitive with entropy-coded quantization.

The user symbols are detected using linear detection,

$$\hat{\mathbf{x}} = \sum_{l=1}^{L} \mathbf{W}_l \tilde{\mathbf{z}}_l, \tag{5.78}$$

where $\mathbf{W}_l$ are the MMSE detection matrices

$$\mathbf{W}_l = \mathbf{C}_e \mathbf{G}_l^\dagger \big(\mathbf{I}_N + \boldsymbol{\Phi}_l\big)^{-1}, \tag{5.79}$$

with $\mathbf{C}_e$ the MMSE error covariance matrix,

$$\mathbf{C}_e = \rho\Big(\mathbf{I}_K + \rho\sum_{l=1}^{L} \mathbf{G}_l^\dagger \big(\mathbf{I}_N + \boldsymbol{\Phi}_l\big)^{-1}\mathbf{G}_l\Big)^{-1}, \tag{5.80}$$

and user symbol mean squared errors $e_k = \big[\mathbf{C}_e\big]_{k,k}$.

Practical systems use QAM constellations rather than Gaussian symbols, leading to a quantizer mismatch when the Lloyd-Max quantizers are designed for Gaussian sources (see Section 4.4.6). However, the outlined quantization noise model remains a good approximation, as veri-

[6]Here the quantization noise components are modelled as being uncorrelated, i.e. diagonal $\boldsymbol{\Phi}_l$. Whilst this is true under the Gaussian vector test channel model, it is not strictly true for fixed-rate quantization – consider the case where two near-identical sources are quantized at low resolution. However, at higher resolution, the quantization intervals become very small and the quantization noise correlation becomes negligible for non-identical sources. For example, the work in [140] approximates the quantization noise correlation as decaying with 2^{-2b}, whilst the correlation noise power decays with 2^{-b}. Much prior work, e.g. [163], [123], therefore negates quantization noise correlation.

fied in Figure 5.17, which compares the analytical user symbol mean squared error expression to numerical simulations using both Gaussian and 64QAM symbol alphabets. The analytical

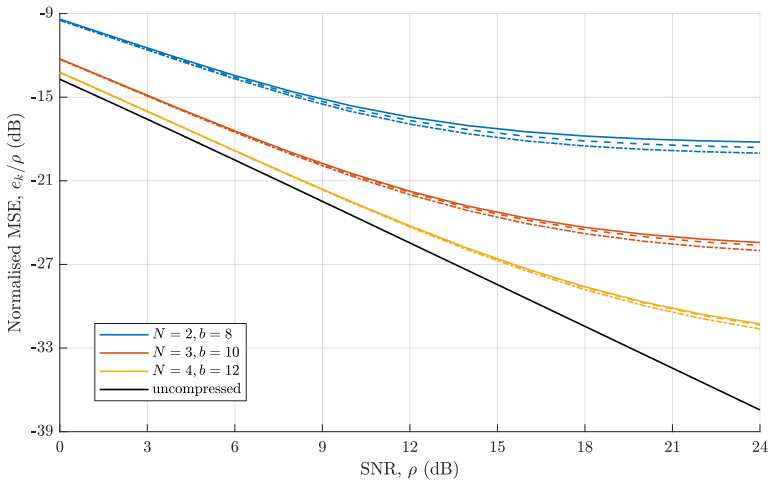

Figure 5.17: User symbol mean squared error after MMSE detection, T-CKLT dimension reduction, fixed-rate scalar quantizers, $K = 8, M = 8, L = 4$. Solid line: analytical expression, dashed line: Gaussian symbols, dot-dash line: 64QAM symbols.

expression approximates the simulated MSE for both Gaussian and QAM symbols, becoming tighter at higher quantizer resolutions. The mean squared error behaves as expected, initially decreasing with SNR before becoming limited by quantization noise and approaching an error floor.

User Capacities

The SINR for user k is given by,

$$\mathrm{SINR}_k = \frac{\rho}{e_k} - 1, \tag{5.81}$$

and the user capacities

$$\mathcal{C}_k = \log_2\left(1 + \mathrm{SINR}_k\right) \tag{5.82}$$

are achievable – since treating the non-Gaussian quantization noise as Gaussian provides a capacity lower bound [196]. Figure 5.18 shows the user mean and 10 % outage capacities for T-CKLT and MF-GS dimension reduction filters and varying signal dimensions and quantizer resolutions. At low fronthaul capacities, a small signal dimension gives the best mean capacities, but will tend to perform poorly in terms of outage capacity. This can be explained by reference to Figure 5.8, since the dimension reduction filters will tend to produce a larger eigenvalue spread when N is small.

Benefit of Additional Antennas

Increasing the number of antennas at each receiver improves the per-user mean and 10% outage capacities, as shown in Figure 5.19. This is consistent with the results from Section 5.3.2, which showed that for fixed N the mutual information and eigenvalue spread of the reduced dimension channel increase with M.

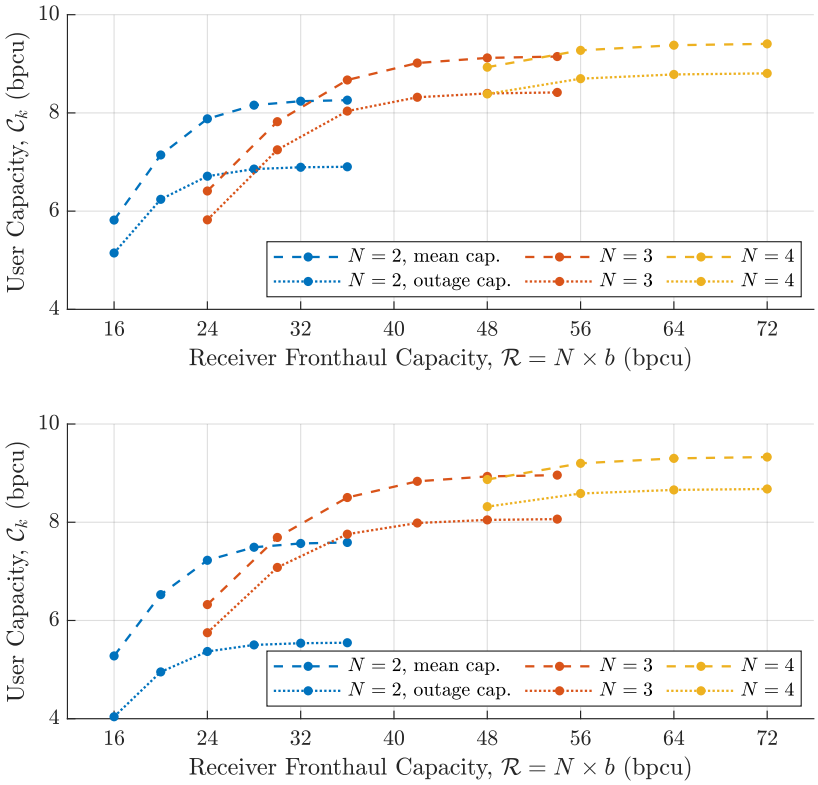

Figure 5.18: User mean and 10 % outage capacities under fixed-rate scalar quantization with varying signal dimensions and quantizer resolutions, $\rho = 15$ dB, $K = 8, M = 8, L = 4$. Top: T-CKLT dimension reduction filters, bottom: MF-GS dimension reduction filters.

Signal Compression under Imperfect CSI

Using the dimension reduction filters for imperfect CSI in Section 5.3.4, the fixed-rate quantization method can be readily extended for the case of imperfect CSI at the receivers. Under the same assumptions for quantizing signals with imperfect CSI as used in Chapter 4, the MMSE estimation error (averaged over possible error realisations) is

$$\mathbf{C}_e = \mathbb{E}_{\mathbf{E}}\big[(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^\dagger\big] = \rho\Big(\mathbf{I}_K + \rho \sum_{l=1}^{L} \hat{\mathbf{G}}_l^\dagger \big(\mathbf{I}_N + \mathbf{\Phi}_l\big)^{-1} \hat{\mathbf{G}}_l\Big)^{-1}, \tag{5.83}$$

where the quantization noise power for dimension i at receiver l is

$$\mathbb{E}\big[|\phi_{l,i}|^2\big] = \rho \mathbf{a}_{l,i}^\dagger \mathbf{\Omega}^{-1/2} \mathbf{H}\mathbf{H}^\dagger \mathbf{\Omega}^{-1/2} \mathbf{a}_{l,i} + \mathbf{a}_{l,i}^\dagger \mathbf{\Omega}^{-1} \mathbf{a}_{l,i}. \tag{5.84}$$

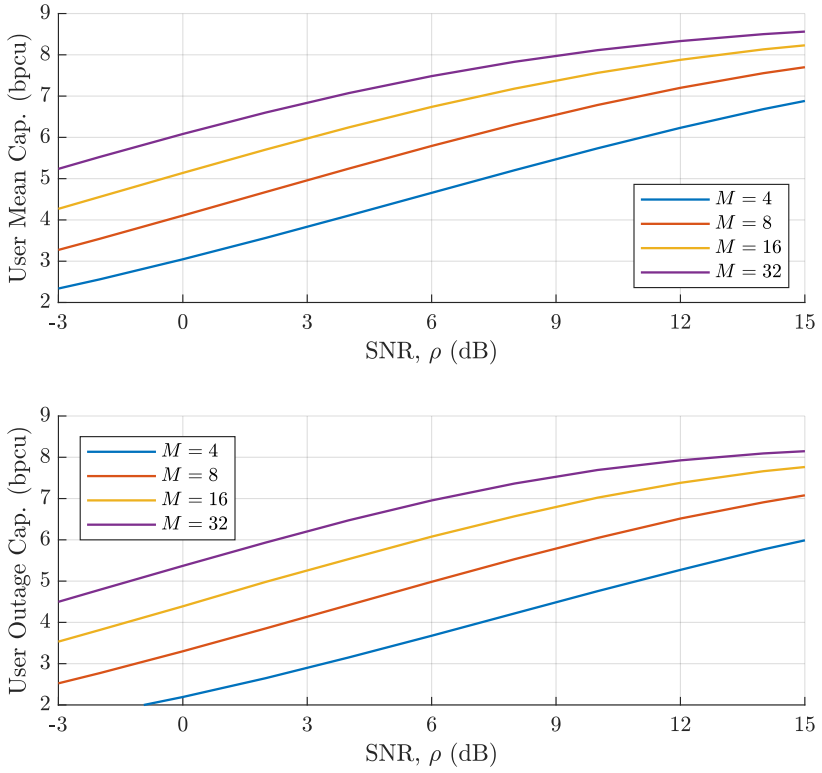

Figure 5.19: User mean and outage capacities under fixed-rate scalar quantization with MF-GS dimension reduction filters, $K = 8, L = 4, N = 3, b = 10$ bits ($\mathcal{R} = 30$ bpcu).

5.4.3 Example: Dense Deployment

Inspired by the cell-free MIMO concept, an example use of dimension reduction-based compression in a dense distributed uplink MIMO C-RAN deployment is now considered. Here, $K = 16$ users are simultaneously served by $L = 16$ receivers, all randomly distributed within a 200 m $\times$ 200 m service area[7] – representing, for example, a dense urban deployment. Each receiver applies MF-GS dimension reductionfilters to its multi-antenna received signal to produce $N = 2$ signal observations[8] (per subcarrier). These observations are then quantized using b-bit fixed-rate quantizers, such that each receiver uses $\mathcal{R} = 2b$ bpcu of fronthaul capacity.

The system operates in a 60 MHz bandwidth, with spectral efficiency

$$\text{mean user spectral efficiency} = 0.8 \times \text{mean user capacity}, \tag{5.85}$$

with the 20 % loss (compared to ideal Nyquist rate signalling) used to account for signalling overheads.

Figure 5.20 shows the capacity for different qualities of channel estimation, assuming each receiver has $M = 4$ antennas and uses $b = 8$ bit quantizers. This corresponds to a fronthaul load of 960 Mbps per receiver – well within what can be provided by mmWave point-to-point

[7] Equivalent, on average, to deploying a receiver every 50 m.

[8] In practice, the F-MF-GS scheme, which gives similar performance, could be used instead to reduce signalling overheads

links.

The benefits of having accurate CSI are clear – with poor quality CSI the capacity is severely limited due to channel estimation errors, whilst with higher quality CSI the effects of quantization noise become more significant (resulting in better fronthaul utilisation). At higher SNRs, mean user throughputs of 300 Mbps can be achieved, corresponding to a total cell throughput of ~ 5 Gbps within the 60 MHz channel.

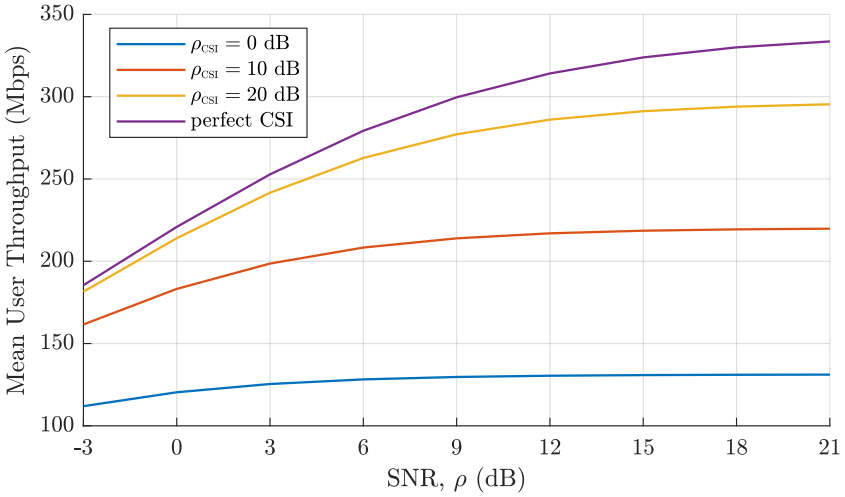

Figure 5.20: Mean user throughputs in dense distributed MIMO deployment with MF-GS dimension reduction and different channel estimation qualities, $N = 2$, $b = 8$ bits, $K = 16, L = 16, M = 4$.

Figure 5.21 shows the effects of increasing the quantizer resolutions and of increasing the number of antennas at each BS, in a system with good quality CSI ($\rho_{\text{CSI}} = 20$ dB). At low SNR, the effect of the quantizer resolution is minimal, since the system is primarily noise-limited, whereas at high SNR the system becomes fronthaul-limited and a substantial increase in throughput can be gained from adding more fronthaul capacity. The array gain provided by increasing the number of receiver antennas can be used to decrease the user transmit power in the noise-limited region (low SNR), or to increase user throughput at high SNR.

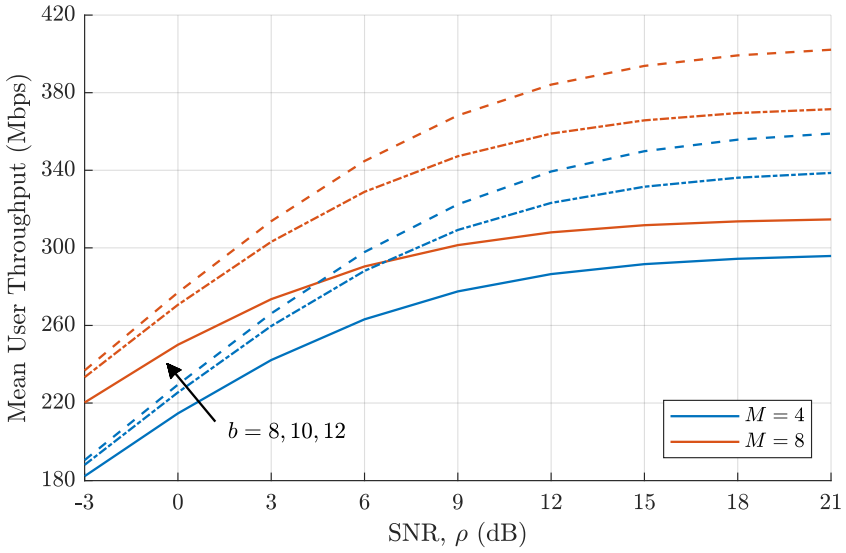

Figure 5.21: Mean user throughputs in dense distributed MIMO deployment with MF-GS dimension reduction, $\rho_{\text{CSI}} = 20$ dB, $N = 2, K = 16, L = 16$. Solid line: per-receiver fronthaul throughput 960 Mbps, dot-dash line: fronthaul throughput 1200 Mbps, dashed line: fronthaul throughput 1440 Mbps.

5.5 Two-Stage Reduced Dimension Precoding for the Distributed MIMO Downlink

So far in this thesis the fronthaul compression efforts have focused on the uplink. However, exploiting the duality between MIMO uplink and downlink, the dimension reduction method is straightforward to adapt for the distributed MIMO downlink – establishing it as a general signalling approach for networks with an excess of antennas and limited fronthaul.

On the downlink, dimension reduction can be applied using two-stage precoding, following the process of the uplink compression method in reverse:

- In the first (inner) precoding stage, the CP precodes the user symbols to produce L reduced dimension signals with $N < M$ dimensions

$$\mathbf{z}_l = \mathbf{W}_l \mathbf{P}^{1/2} \mathbf{s}. \tag{5.86}$$

- The reduced dimension precoded signals are then quantized using N fixed-rate scalar quantizers for transfer over finite-capacity fronthaul

$$\tilde{\mathbf{z}}_l = \mathbf{z}_l + \boldsymbol{\delta}_l \tag{5.87}$$

- In the second (outer) precoding stage, each transmitter beamforms its compressed signal from M antennas using N beamforming vectors

$$\mathbf{x}_l = \mathbf{A}_l \tilde{\mathbf{z}}_l. \tag{5.88}$$

Each user then receives the simultaneous transmissions from all L transmitters

$$y_k = \sum_{l=1}^{L} \mathbf{h}_{l,k}^T \mathbf{A}_l \big(\mathbf{W}_l \mathbf{P}^{1/2} \mathbf{s} + \boldsymbol{\delta}_l \big) + \eta \tag{5.89}$$

from which it decodes its intended symbol.

The proposed precoding has parallels to the sparse precoding approach, in that each of the transmitters only transmits into a reduced subspace of its available channel space. However, unlike with sparse precoding – where each transmitter explicitly transmits data to only a subset of the active users – here each transmitter serves all users with its precoded signal (but may transmit more power towards certain users due to the precoding).

5.5.1 Two-Stage Precoder Design

For tractability, here the outer precoders are designed first, followed by the inner precoders. The effects of quantization noise are then accounted for, and power control schemes investigated.

Outer Precoder Design

The outer precoding stage takes the N-dimension quantizated signal produced by the inner precoder and beamforms it from M antennas. If the N beamforming vectors at receiver l are constrained to be orthonormal then the outer precoding matrix $\mathbf{A}_l$ is semi-orthogonal, and preserves the power of the reduced dimension signal

$$P_T = \mathbb{E}\big[\mathbf{x}_l^\dagger \mathbf{x}_l\big] \tag{5.90}$$

$$= \mathbb{E}\big[\tilde{\mathbf{z}}_l^\dagger \mathbf{A}_l^\dagger \mathbf{A}_l \tilde{\mathbf{z}}_l\big] = \mathbb{E}\big[\tilde{\mathbf{z}}_l^\dagger \tilde{\mathbf{z}}_l\big]. \tag{5.91}$$

The inner precoder can then be thought of acting on a set of L reduced dimension channels,

$$\mathbf{G}_l = \mathbf{H}_l \mathbf{A}_l. \tag{5.92}$$

The $\mathbf{A}_l$ may then be chosen to maximise the capacity of an equivalent reduced dimension channel that exists between the $\tilde{\mathbf{z}}_l$ and received signals, $\mathbf{y}$. For simplicity, it is assumed here that the distributed MIMO network is subject to a network power constraint of LP_T (rather than individual transmitter power constraints, P_T), shared equally between the user streams. Then by uplink-downlink duality,

$$\mathcal{C}_{\text{EQ}} = \log_2 \det \Big(\mathbf{I}_K + \frac{LP_T}{K} \sum_{l=1}^{L} \mathbf{H}_l \mathbf{A}_l \mathbf{A}_l^\dagger \mathbf{H}_l^\dagger \Big). \tag{5.93}$$

Comparing to (5.11), outer precoding matrices can be found in an identical manner to the uplink dimension reduction filters, using the T-CKLT or MF-GS methods – but applied in *reverse* – to produce reduced dimension channels with the same properties. This is attractive from an implementation perspective when the same users are being served by a TDD system on both uplink and downlink, since the uplink dimension reduction filters can potentially be re-used for

downlink precoding, reducing signalling overheads and computation.

Inner Precoder Design

Since the outer precoding stage at each transmitter preserves the power of its signal, the inner precoders can simply be designed to act on the equivalent reduced dimension channels, $\mathbf{G}_l$. Here for simplicity the effects of quantization are ignored whilst initially designing the inner precoding stage.

Restricting attention to ZF precoding, which eliminates all inter-user interference and is optimal at high SNR, there are two possible strategies:

1. Using the Moore-Penrose pseudo-inverse, with

$$\mathbf{W}_l = \mathbf{G}_l^{\dagger}\big(\sum_{j=1}^{L}\mathbf{G}_j\mathbf{G}_j^{\dagger}\big). \tag{5.94}$$

   To meet per-transmitter power constraints, power control must then be performed. This method is attractive from the perspective of computational complexity, but is sub-optimal and can result in some transmitters transmitting at well below their power constraint.

2. Choose the precoding matrices under multiple per-transmitter power constraints (ZF-MPC). This is more computationally expensive since no closed-form solutions exist and the precoding matrices must therefore be found numerically as described in Section 2.5.2. However this method better utilises the power available in the network, for improved user performance[9].

The outputs of the precoders are then quantized using N fixed-rate scalar quantizers with resolution b bits, producing

$$\tilde{\mathbf{z}}_l = \mathbf{z}_l + \boldsymbol{\delta}_l \tag{5.95}$$

where $\mathbb{E}\big[\boldsymbol{\delta}_l\boldsymbol{\delta}_l^{\dagger}\big] = \boldsymbol{\Phi}_l$, with quantization noise power on component i

$$\phi_{l,i} = \big[\boldsymbol{\Phi}_l\big]_{i,i} \approx \mathbf{e}_i^{\dagger}\mathbf{W}_l\mathbf{P}\mathbf{W}_l\mathbf{e}_i\frac{\sqrt{3}\pi}{2}2^{-b}, \tag{5.96}$$

where $\mathbf{e}_i$ is the unit vector consisting of $(N-1)$ zeros with a single one in position i. The received signal is

$$y_k = \sum_{l=1}^{L}\mathbf{g}_{l,k}^{T}\big(\mathbf{W}_l\mathbf{P}^{1/2}\mathbf{s} + \boldsymbol{\delta}_l\big) + \eta, \tag{5.97}$$

where $\mathbf{g}_{l,k} = \mathbf{A}_l^{\dagger}\mathbf{h}_{l,k}$ is the reduced dimension channel between user k and transmitter l, and

[9]Note that when $NL = K$, the matrix inverse is unique and the standard ZF and ZF-MPC matrices are therefore identical.

the SINR at user k is given by

$$\mathrm{SINR}_k = \frac{\rho_k \left| \sum_{l=1}^{L} \mathbf{g}_{l,k}^T \mathbf{w}_{l,k} \right|^2}{\sum_{l=1}^{L} \mathbf{g}_{l,k}^T \mathbf{\Phi}_l \mathbf{g}_{l,k} + 1}. \tag{5.98}$$

Here, for space reasons, only the perfect CSI case is considered, but this analysis is readily extended to include the effects of both quantization noise and imperfect CSI by incorporating CSI errors into the SINR as outlined in Section 2.4.2.

Figure 5.22 shows the mean capacity achieved using standard Moore-Penrose ZF, and for inner ZF precoders designed under per-transmitter power constraints (ZF-MPC). In both scenarios, power control is applied such that all users have the same received signal strength, with a per-transmitter power constraint of 24 dBm in a 20 MHz channel – the 3GPP-defined transmit power of a 'local area' small cell [76]. For comparison, conventional (single-stage) ZF-MPC precoding with quantization is also shown, where the number of antennas at each transmitter is reduced so that the fronthaul load is the same as under two-stage precoding case.

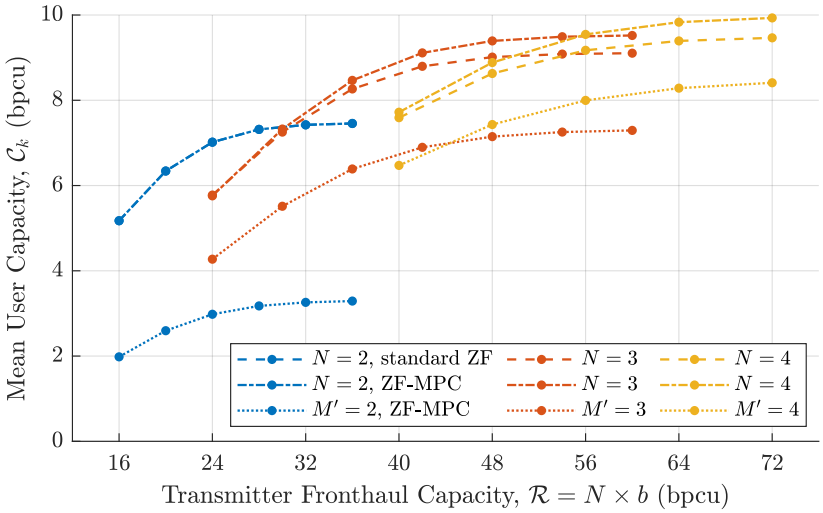

Figure 5.22: Mean downlink user capacities under two-stage precoding with T-CKLT outer precoder design and ZF/ZF-MPC inner precoder design. $P_T = 24$ dBm, $B = 20$ MHz, receiver noise figure 5 dB, $K = 8, L = 4, M = 8$.

Overall, similar behaviour is seen to the uplink compression scheme in Figure 5.18 – increasing the quantizer resolution increases the user capacity until capacity levels off due to loss from dimension reduction, at which point increasing N is beneficial. The two-stage precoding significantly outperforms single-stage precoding, whilst, here, the benefit of using the more complex inner precoder design is relatively small.

5.5.2 Quantization-aware Power Allocation

Designing the inner ZF precoder to produce the same received signal strength at all users does not result in equal user SINRs, since the received user signals contain different amount

of quantization noise. Since the received quantization noise depends on the user powers, as shown in (5.96) & (5.98), the user power allocations can be chosen to account for the effect of quantization noise and improve performance – an idea previously applied in an uplink context in [12] and [123].

Following some tedious manipulation, it is straightforward to show that the SINR of each user can be expressed in the form

$$\mathrm{SINR}_k = \frac{\rho_k}{a_k + \sum_{j=1}^{K} b_{k,j}\rho_j}. \tag{5.99}$$

From this it can be seen that the SINR of any user can be improved by reducing the powers allocated to the other users – at the expense of reducing their SINRs. Max-min power allocation can therefore be used to improve the performance of the users worst affected by quantization noise, solving

$$\begin{aligned} &\underset{\rho_1,\ldots,\rho_K}{\text{maximise}} \quad \min_k \mathrm{SINR}_k \\ &\text{subject to} \quad \mathrm{Tr}\left(\mathbf{W}_l \mathbf{P} \mathbf{W}_l^{\dagger}\right) \leq P_T \quad \forall l. \end{aligned} \tag{5.100}$$

This can then be solved using standard geometric programming as described in Section 2.4.3.

The max-min power allocation results in all users receiving an equal SINR – increasing the SINR of the worst users whilst decreasing that of the best users. This has the effect of reducing the variability of the user capacities, as shown in Figure 5.23 for both standard ZF (top) and ZF-MPC (bottom) precoders.

At lower quantizer resolutions, where user capacity is limited by quantization noise, max-min power control significantly improves the performance of the worst users, with little variation in user capacities. Comparing the top and bottom figures, here the use of max-min power control has a much bigger impact on performance than the choice of precoder – indicating that when computational resources are limited it is better to use them for power control than for precoder optimisation. As the quantizer resolution increases, the impact of quantization noise reduces, and max-min power control has a much smaller impact.

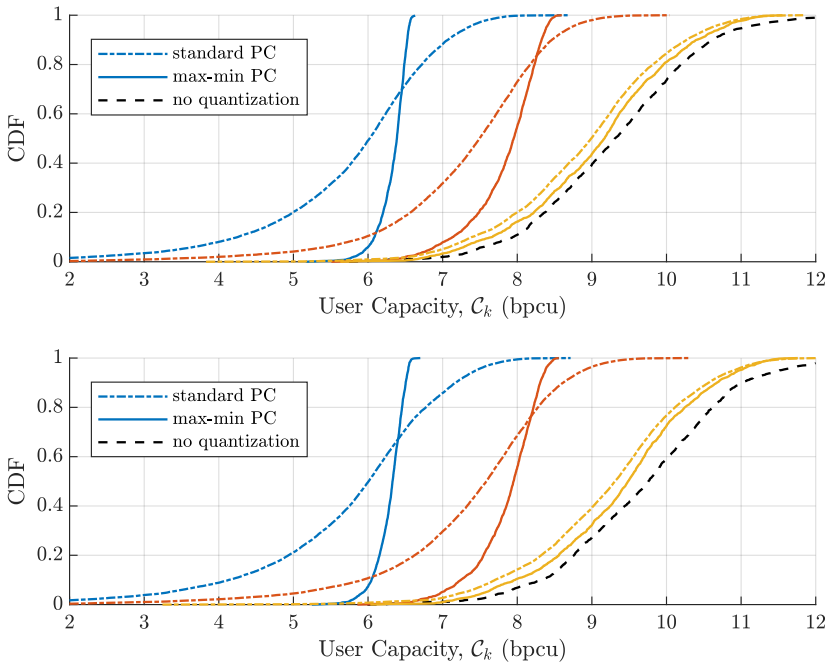

Figure 5.23: Performance of two-stage precoding with and without quantization-aware max-min power control, MF-GS outer precoder design, $P_T = 24$ dBm, $B = 20$ MHz, noise figure 5 dB, $K = 8, L = 4, M = 8, N = 3$. Top: standard ZF inner precoder, bottom: ZF-MPC inner precoder design. Blue line: $b = 8$ bpcu, orange line: $b = 10$ bpcu, yellow line: $b = 12$ bpcu.

5.5.3 Example: Dense Deployment

Numerical mean user throughput results are now presented for the same dense distributed MIMO example considered in Section 5.4.3. In this example, MF-GS outer precoder design is used in conjunction with the lower-complexity standard ZF inner precoding and quantization-aware max-min power control. For reference, conventional single-stage precoding using a reduced number of antennas is also shown. Perfect CSI is assumed in all cases. The mean user throughputs within the 60 MHz channel are shown in Figure 5.24.

The use of two-stage precoding provides a mean throughput benefit compared to conventional single-stage precoding of around 40 Mbps per user – 640 Mbps of total additional throughput. At low transmit power, the use of additional transmit antennas increases the throughput, but these gains reduce at higher power as quantization noise becomes the limiting factor[10].

Using $b = 8$ bit quantizers and practical transmit powers (e.g. 25 dBm), 16 transmitters each equipped with 1 Gbps of fronthaul capacity are capable of providing the 16 users with an average throughput of 320 Mbps within the 60 MHz channel – representing an overall fronthaul efficiency of around 35%. When 1.5 Gbps of fronthaul capacity is available a mean user throughput of 480 Mbps can be achieved, for a total throughput of over 7.5 Gbps.

[10]The diversity/reliability benefits of increasing the number of transmit antennas are not shown in this figure.

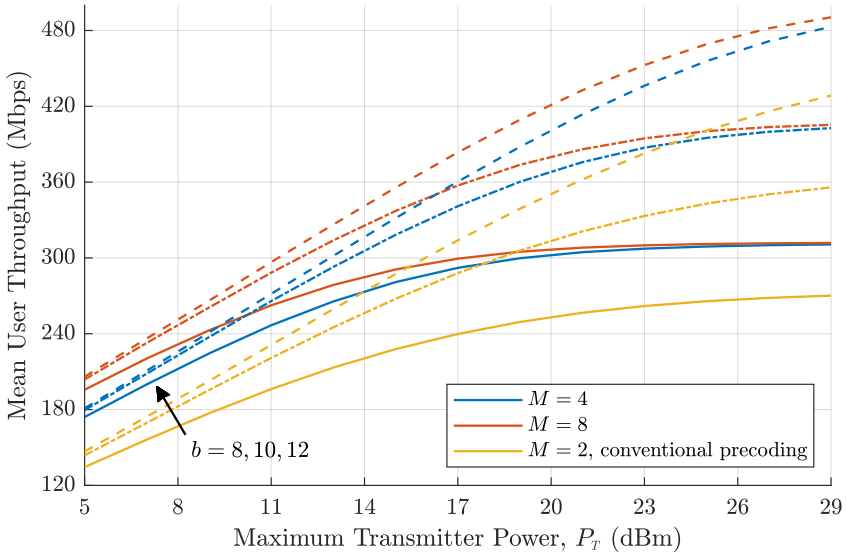

Figure 5.24: Mean user downlink throughputs in dense distributed MIMO deployment with MF-GS outer precoders, $B = 60$ MHz, receiver noise figure 5 dB, $N = 2, K = 16, L = 16$. Solid line: per-transmitter fronthaul throughput 960 Mbps, dot-dash line: fronthaul throughput 1200 Mbps, dashed line: fronthaul throughput 1440 Mbps.

5.6 Conclusion

When the remote radio heads in a distributed MIMO network are connected to the CP via finite-capacity fronthaul links, exploiting the benefits of deploying an excess of antennas is challenging due to the corresponding increase in fronthaul data. This chapter investigates the use of distributed dimension reduction to reduce the amount of data that must be transferred over fronthaul whilst trying to preserve the benefits of having additional antennas – an area that has previously received little research attention.

Here, linear dimension reduction is applied across the multiple antennas at each of the radio heads. This combination of MIMO channel and dimension reduction at each radio head can be described by an equivalent reduced dimension MIMO channel, reducing the dimensionality of the distributed MIMO system seen by the CP – and of the signals that must be transferred over fronthaul. This approach can be used both on the uplink – by applying a dimension reduction filter to each received signal – and on the downlink – by beamforming reduced dimension signals using a larger number of antennas – making it a bidirectional approach to fronthaul data reduction.

Distributed Dimension Reduction

The first part of the chapter addresses the problem of finding uplink dimension reduction filters that best preserve the benefits of having additional antennas, to produce a reduced dimension MIMO system with similar properties to the full system. It is shown that on the uplink the dimension reduction filters that jointly capture the maximum information about the uplink symbols are a truncated form of the conditional KLT (T-CKLT), found using an iterative block

co-ordinate ascent procedure. A combination of analysis and numerical results are then used to demonstrate that with these filters, it is possible to significantly reduce the dimensionality of the MIMO system without sacrificing much performance, such that using a large number of antennas is beneficial even when the channel dimension is fixed.

From a practical implementation perspective, the downside of this optimal filter design method is that the filters must be jointly calculated at the CP using full CSI at every coherence interval – resulting in significant fronthaul signalling overheads. An alternative approach is therefore also proposed, in which the dimension reduction filter for each receiver is instead calculated using a defined subset of the local user matched filtering vectors. In fading channels, these filters can be updated locally at the receiver for each coherence interval, with only the reduced dimension CSI required at the CP for symbol detection – reducing signalling overheads at the cost of some loss in performance. Both dimension reduction methods are shown to be straightforward to adapt for the case of imperfect CSI.

As a relatively understudied area, the findings of this chapter raise more questions and research opportunities than there is scope here to address. For example, here the dimension reduction filters are designed under a joint mutual information criteria – which has a clear connection to sum capacity – but many other criteria are possible, perhaps aiming to improve performance under more specific user quality of service metrics. Practical dimension reduction schemes like the F-MF-GS scheme that can achieve good performance with reduced signalling overheads are also an interesting area where there is scope for further development. Different applications of dimension reduction, for example applied across time or frequency blocks as well as antennas, could also be an area of practical research.

From an theoretical perspective, analysing the performance of dimension reduction under specific fading models poses an interesting technical challenge, and may yield more precise insights into the benefits of adding antennas at each receivers and fading diversity providing. Numerical investigations into the performance of dimension reduction using more complex channel models and different system configurations would complement this by providing some more practical insights.

This chapter has also demonstrated that for the downlink, outer precoders/beamformers can be found by simply applying the dimension reduction filters in reverse – exploiting the duality between the uplink and downlink. This enables reduced dimension fronthaul signalling on the downlink using two-stage precoding, and establishes the bidirectionality of the dimension reduction approach. Under TDD operation, the same filters can then be used on both uplink and downlink, reducing signalling overheads. Further work here could investigate other outer precoding design strategies, for example explicitly optimising them under per-transmitter power constraints (rather than the simpler total power constraint used here).

Dimension Reduction-based Signal Compression

This chapter also considers the use of dimension reduction in conjunction with lossy data compression, to enable specific fronthaul capacity constraints to be met. On the uplink, a simple & practical scheme is proposed in which the outputs of the reduced dimension channels are compressed using a set of scalar quantizers with equal rate allocations, and forwarded to the CP

for joint symbol detection. On the downlink, a two-stage precoding is proposed where the inner precoder generates low dimension signals, which are quantized and transferred over fronthaul to the remote transmitters, where the outer precoders are applied.

A theoretical justification for dimension reduction-based signal compression is provided by showing that, at high SNR, the distributed MIMO sum capacity in the quantization noise-limited region scales approximately linearly with the fronthaul capacity, and inversely with the signal dimension. In this region, choosing a small signal dimension enables this simple signal compression scheme to achieve very efficient fronthaul utilisation – coming within a fixed gap of the cut-set bound. At higher fronthaul rates, the quantization noise reduces and it is beneficial to increase the signal dimension.

Numerical examples show this dimension reduction-based signal compression can achieve similar – and often better – performance than the optimal transform coding scheme and the joint rate allocation scheme outlined in Chapter 4. Whilst this initially appears somewhat surprising, there is actually a strong parallel between the joint dimension reduction scheme and the joint rate allocation scheme – the former exploits the sparsity in the network to produce a reduced dimension compressed signal representation, whilst the latter tends to produce a sparse rate allocation, resulting in the compressed signal also having a reduced dimension.

An alternative low complexity uplink compression scheme is proposed that uses matched filter-based dimension reduction in conjunction with low complexity fixed-rate scalar quantizers (instead of entropy-coded scalar quantizers). Numerical results are provided for a dense deployment with 16 users and 16 remote receivers operating under imperfect CSI in a 60 MHz channel, where mean user throughputs of over 300 Mbps are achieved when each receiver has a fronthaul capacity of 1 Gbps, and over 400 Mbps with 1.5 Gbps fronthaul capacity.

This is extended to the downlink using two-stage precoding, where the inner precoding stage uses a ZF precoder matched to the reduced dimension channels, quantization applied prior to the fronthaul. It is shown that when operating in the fronthaul-limited region the use of user max-min power allocation is very effective for mitigating the effects of quantization noise, considerably improving the capacities of the worst users. Numerical results for the dense deployment show that similar user throughputs can be achieved on the downlink as on the uplink.

The findings presented in this chapter suggest that dimension reduction is a promising low complexity bidirectional fronthaul data compress/reduction strategy. Future work should look to build on these results and investigate the use of dimension reduction in a range of different propagation environments and system configurations. This should include a thorough investigation into other important practical aspects such as the signalling overheads and latency constraints, the latter being particularly important in downlink networks.

Chapter 6

Conclusion

Multi-user MIMO technology is set to play an important role in enabling fifth generation and future wireless systems to achieve the high spectral efficiencies and area capacities required to meet the growing demand for mobile data services. The past decade has seen the development of two particular architectures – massive MIMO & distributed MIMO C-RAN – that have the potential to unlock the benefits of MU-MIMO on a much larger scale than previously seen in cellular systems. This PhD, and the research outlined in this thesis, has contributed to the development of scalable & energy efficient commercial MU-MIMO systems by focusing on two distinct practical challenges associated these architectures: PAPR reduction for the massive MIMO downlink, and fronthaul data compression/reduction for distributed MIMO C-RAN systems.

The research outlined in this thesis has focused on the development of novel signal processing-based solutions to the above problems, with an emphasis on low complexity practical methods that are appropriate for use in commercial systems. Along with the proposed solutions, this research has uncovered a variety of useful insights and potential directions for future work. A brief summary of these is now provided.

PAPR Reduction for Massive MIMO

The high peak-to-average power ratio of precoded MIMO-OFDM downlink signals necessitates the use of a large power backoff at the transmitting power amplifiers, reducing energy efficiency and increasing the amplifier peak power requirements. Chapter 3 focuses on the challenge of PAPR reduction for the massive MIMO downlink, aiming to reduce the PAPR of the transmitted signals with minimal impact on link performance.

A clipping-based approach to PAPR reduction is investigated, producing a number of novel technical contributions, as listed in **Section 3.1.1**. The key findings & insights from Chapter 3 can be summarised as:

1. The distortion introduced when precoded MIMO-OFDM signals are clipped causes two effects. Previous analysis has generally used a simple additive error model to describe the effects of clipping in MIMO systems, but in Section 3.3 a more rigorous statistical model for clipping based on Bussgang's theory is developed. This model shows that clipping

distortion manifests itself in two effects: it attenuates & distorts the MIMO precoding to introduce additional inter-user interference at the receivers, and it introduces random clipping noise at the receivers. Although analysing the exact dynamics of this model is challenging due to its non-linearity, this model provides some simple insights – the clipping noise will tend to be correlated across antennas (rather than independent as often assumed), precoding distortion will always be produced when the per-antenna average transmit power varies, and clipping distortion can produce a near-far effect where users close to the BS suffer more significant performance degradation due to the presence of more distant users.

2. Iterative clipping & spatial filtering is an effective & practical strategy for PAPR reduction, and should be designed to account for both clipping effects. Sections 3.5.2 & 3.5.3 develop a novel iterative clipping & spatial filtering scheme that exploits the large massive MIMO nullspace to construct a least squares approximation of the clipped signal that eliminates the clipping distortion from the received user signals. Whilst related schemes have previously been proposed, this scheme incorporates the Bussgang model to properly account for the attenuation & distortion of the MIMO precoding, enabling higher levels of PAPR reduction to be achieved without sacrificing link performance. Numerical results show that over 8 dB of PAPR reduction can be achieved, 1 dB more than comparable schemes.

3. Active constellation extension can improve the PAPR reduction achieved by clipping & spatial filtering. Section 3.5.5 shows that the proposed iterative clipping & spatial filtering scheme can easily be extended to include active constellation extension, enabling greater levels of PAPR reduction to be achieved. Numerical results demonstrate that this is particularly useful when smaller constellation sizes are used – providing up to 1 dB additional PAPR reduction – or in smaller scale MIMO systems with fewer BS antennas – where over 2 dB additional PAPR reduction can be achieved.

4. The MIMO channel & precoding have a significant impact on the signal PAPR. Numerical results provided in Section 3.2.3 show that under a correlated fading channel model the PAPR of the precoded MIMO-OFDM signal is 2+ dB higher than under i.i.d fading. This appears to be at least partially down to variation in average transmit powers across the array, which necessarily increases PAPR and can be addressed by precoding under per-antenna power constraints. Numerical results in Section 3.5.4 show that the proposed solution achieves similar PAPR *reduction* in both scenarios, but the final PAPR under correlated fading remains higher than with i.i.d fading (under conventional symbol precoding).

Directions for Future Work

The findings of Chapter 3 point to two interesting and useful directions for future investigation. The first is assessing the benefits of the proposed solution in a broader range of propagation environments and system configurations. Whilst the scheme is very effective at reducing PAPR whilst maintaining link performance at higher capacities, it is known that operating at low SNR or with poor quality CSI the relative impact of clipping noise on performance is reduced, and

hence the use of conventional clipping & filtering without the spatial filtering may be sufficient in these environments.

Further work to thoroughly investigate the impact of the precoding on PAPR in different propagation environments is the other key area for future work. This should include a thorough investigation into the influence of per-antenna power variations on PAPR, and could involve the development of low complexity precoding schemes to reduce these power variations – complemening the proposed solution to bring further PAPR reduction.

Signal Compression for Uplink Massive MIMO C-RAN

The cloud radio access network (C-RAN) architecture uses a shared central processor to perform the baseband signal processing & network functions for multiple radio units, decreasing cost of deployment & operation and increasing network flexibility. However, when the fronthaul connections between the radio unit and CP have limited capacity – for example with ethernet or wireless point-to-point links – directly transferring the raw sampled data produced by the large number of antennas on the massive MIMO uplink rapidly becomes infeasible. The first part of Chapter 4, Section 4.3, investigates the application of signal compression to the received uplink signals, aiming to maximise performance under fronthaul capacity constraints.

The use of transform coding-based compression of the received uplink signals is considered, producing a number of novel results, as listed in **Section 4.1.1**. The key findings & insights from Section 4.3 can be summarised as:

1. Transform coding can exploit the underlying sparsity of uplink massive MIMO signals to achieve efficient signal compression. The large excess of antennas used in a massive MIMO system causes the received signals to have an inherent underlying sparsity. Section 4.3.2 shows that the Karhunen-Loeve transform can exploit this sparsity to reduce the signal dimension. Section 4.3.2 then shows that applying lossy compression to the transformed signals using a set of optimal scalar lossy compressors with appropriate rate (bit) allocations asymptotically achieve the cut-set bound at high SNR – efficiently utilising the available fronthaul. This means that when operating in the fronthaul-limited region, increasing the available fronthaul capacity by a certain quantity increases the achievable sum capacity by a similar amount – transform coding therefore represents an effective strategy for reducing the amount of sampled data for transfer over fronthaul.

2. Transform coding can capture many of the benefits of deploying excess antennas, even when capacity is fundamentally limited by the fronthaul Section 4.3.4 investigates what the benefits (if any) are of deploying massive MIMO in scenarios where the MIMO capacity is fundamentally limited by the capacity of the fronthaul link. It is shown using a combination of analysis and numerical simulations that under the proposed transform coding scheme many of the well known benefits of deploying a large number of antennas are still seen – the effects of fast fading disappear, transmit power reduces and linear detection methods become optimal.

Directions for Future Work

The findings of Chapter 4 demonstrate that transform coding is an effective and practical method for compressing the large amount of received signal data produced by a massive MIMO receiver, making it feasible to deploy in situations where only limited capacity fronthaul connections can be provided. However, these results are based on analysis that assumes the use of Gaussian transmit symbols and an ideal lossy compression model, which represent approximations of the actual constellation symbols and entropy-coded quantization methods used in practical systems – which should be tested & validated using bit level simulations in future work.

At a more fundamental level, it is known that with a single receiver the use of signal compression is a sub-optimal strategy – perfect fronthaul utilisation can be achieved by instead performing symbol detection at the remote receiver and transferring the decoded data streams back over fronthaul. Transform coding achieves lower capacity compared to this detect & forward case, but may still be an attractive alternative when a functional split with less processing at the remote receiver and more complexity at the CP is desired. This is an important area for further investigation, with signalling overheads and energy efficiency being key considerations.

Data Compression/Reduction for Distributed MIMO C-RAN

The use of low capacity fronthaul connections will play an important role in enabling dense, flexible distributed MIMO C-RAN deployments that can achieve good coverage in areas with high traffic demand. However, the analysis in Section 4.4.1 demonstrates that applying signal compression at each remote receiver by replicating the single receiver transform coding scheme from Section 4.3.2 produces poor performance, because it does not exploit the statistical dependencies between the received signals.

The remainder of Chapter 4 and the whole of Chapter 5 investigate data compression strategies for distributed MIMO. First, Section 4.4 proposes a transform coding solution based on adapting the single receiver scheme from Section 4.3.2, before Chapter 5 investigates a dimension reduction-based approach for distributed networks with an overall excess of antennas. The key technical contributions to the state-of-the-art are listed in **Sections 4.1.1 & 5.1.1**, with the following key findings & insights gained:

1. Transform coding using jointly optimised rate allocations is a scalable approach to achieving efficient data compression on the distributed MIMO uplink. Section 4.4.3 proposes a modification to the transform coding scheme used in Section 4.3.2 where each receiver compresses its transformed signal using scalar compressors with jointly optimised rate allocations. These rate allocations are optimised across all receivers in order to maximise the sum capacity achieved under joint detection, accounting for the inter-receiver signal dependencies. Numerical results demonstrate that the proposed scheme suffers only a small performance penalty compared to the optimal point-to-point compression scheme – which has significantly higher complexity. Numerical examples demonstrate that with 4 remote receivers, each equipped with 8 antennas and 1.5 Gbps fronthaul capacity, a total mean user throughput of over 420 Mbps

could be simultaneously supplied to 8 users within a 100 MHz channel – an improvement of 100 Mbps per user compared to the case where the compression rates are not jointly optimised.

Whilst transform coding with a joint rate allocation has been previously proposed elsewhere, the method proposed here has the advantage of not requiring any intensive numerical solvers – instead using an iterative approach that has closed form solutions at each iteration – and represents a scalable solution for implementing in real time systems.

2. The inherent sparsity of a distributed MIMO network with an overall excess of BS antennas can be effectively exploited using dimension reduction. Similarly to the massive MIMO case, when the distributed MIMO network has an overall excess of BS antennas, the received signals are characterised by inherent sparsity. The findings of Section 5.3 show that distributed dimension reduction can exploit this sparsity – applying a simple dimension reduction filter at each multi-antenna receiver to produce a reduced dimension MIMO system. When these filters are optimally designed to maximise their joint mutual information, as described in Section 5.3.1, the dimensionality of the MIMO system can be significantly reduced without sacrificing much performance. This enables the benefits of using a large number of antennas to be captured whilst the dimensionality of the fronthaul data is restricted. A second low complexity scheme, proposed in Section 5.3.3, exploits these opportunities further to also reduce the CSI that must be transferred over fronthaul by using a set of matched filters to perform dimension reduction at each receiver.

3. Dimension reduction plays an important role in data compression for fronthaul-limited distributed MIMO uplink C-RAN. Section 5.4.1 shows that, at high SNR, applying lossy scalar compression to the outputs of the dimension reduction filters results in a sum capacity that scales approximately linearly with the available fronthaul capacity, and inversely with the signal dimension. Choosing the minimum signal dimension then enables the sum capacity in the fronthaul-limited region to come within a fixed gap of the cut-set bound. Numerical results demonstrate that dimension reduction-based signal compression can achieve similar performance to (and often better than) the joint rate allocation transform coding scheme from Section 4.4.3.

In fact, there is a strong connection between the two proposed schemes, with Section 4.4.3 showing that the jointly optimised rate allocation will tend to be sparse in the fronthaul-limited region – performing an implicit dimension reduction by only compressing a subset of the available signal components. This indicates that dimension reduction plays an intrinsic role in efficient data compression in the fronthaul-limited region.

4. Dimension reduction is an effective & low complexity bidirectional approach to fronthaul data compression/reduction. Finally, Section 5.5.1 extends the dimension reduction approach to the downlink of distributed MIMO systems, using a two-stage precoding approach. Here, duality is exploited to show that the uplink dimension reduction filters can be reversed and used as a downlink outer precoder, with an inner precoder at the CP generating low dimension signals that are transferred over fronthaul. Numerical results provided using simple fixed-rate scalar quantizers for lossy compression show that the dimension reduction

approach can be employed to provide fronthaul data compression on both the uplink (Section 5.4.3) and the downlink (Section 5.5.3) – establishing it as a simple but effective bidirectional approach to fronthaul data compression/reduction.

Directions for Future Work

As an area that has previously received very little research attention, the use of dimension reduction in distributed MIMO networks opens up many areas for potential future research. From a theoretical perspective, characterising the performance of these networks under specific fading models poses an interesting challenge – one beyond the scope of this PhD. The design of alternative dimension reduction schemes is also an area that could be further explored both on the uplink and downlink, and poses many opportunities. For example, other dimension reduction criteria that relate more specifically to user quality of service could be investigated, along with dimension reduction schemes that operate across time or frequency blocks rather than just antennas. Future work should look to build on these results and investigate the use of dimension reduction in a range of different propagation environments and system configurations.

Further investigation into practical aspects of fronthaul signalling are also required – looking at the trade-offs between performance and signalling overheads in mobile channels, and latency constraints, which are particularly important in downlink networks.

Concluding Remarks

This thesis has investigated two distinct challenges associated with implementing MU-MIMO technology on a large scale in cellular systems. Using a combination of theoretical analysis, numerical simulation and pragmatic consideration of realistic constraints, practical solutions to these problems have been proposed that could play a role in future wireless systems.

Appendices

1 Clipping-based PAPR Reduction for the Massive MIMO Downlink

1.1 Least Squares Derivation

Derivation of solution to

$$\underset{\mathbf{x}_{\mathrm{LS}}}{\text{minimise}} \quad \left\|\mathbf{x}_{\mathrm{LS}} - \mathbf{x}_{\mathrm{CF}}\right\|^2, \tag{1}$$

$$\text{subject to} \quad \mathbf{H}\mathbf{x}_{\mathrm{LS}} = \mathbf{H}\mathbf{x}. \tag{2}$$

Making the substitutions

$$\mathbf{a} = \mathbf{x}_{\mathrm{LS}} - \mathbf{x}_{\mathrm{CF}}, \tag{3}$$

$$\mathbf{b} = \mathbf{H}(\mathbf{x} - \mathbf{x}_{\mathrm{CF}}), \tag{4}$$

gives a standard least norm problem

$$\underset{\mathbf{a}}{\text{minimise}} \quad \|\mathbf{a}\|^2, \tag{5}$$

$$\text{subject to} \quad \mathbf{H}\mathbf{a} = \mathbf{b}, \tag{6}$$

which is well known to be solved by the Moore-Penrose pseudo-inverse

$$\mathbf{a} = \mathbf{H}^\dagger\left(\mathbf{H}\mathbf{H}^\dagger\right)^{-1}\mathbf{b}. \tag{7}$$

Re-substitution yields

$$\mathbf{x}_{\mathrm{LS}} = \mathbf{x}_{\mathrm{CF}} - \mathbf{H}^\dagger\left(\mathbf{H}\mathbf{H}^\dagger\right)^{-1}\mathbf{H}(\mathbf{x}_{\mathrm{CF}} - \mathbf{x}). \tag{8}$$

By the same method, the solution to

$$\underset{\mathbf{x}_{\mathrm{BLS}}}{\text{minimise}} \quad \left\|\mathbf{x}_{\mathrm{BLS}} - \mathbf{x}_{\mathrm{CF}}\right\|^2, \tag{9}$$

$$\text{subject to} \quad \mathbf{H}\mathbf{x}_{\mathrm{BLS}} = \mu\mathbf{H}\mathbf{x}, \tag{10}$$

can be shown to be

$$\mathbf{x}_{\mathrm{BLS}} = \mathbf{x}_{\mathrm{CF}} - \mathbf{H}^\dagger\left(\mathbf{H}\mathbf{H}^\dagger\right)^{-1}(\mathbf{H}\mathbf{x}_{\mathrm{CF}} - \mu\mathbf{H}\mathbf{x}). \tag{11}$$

1.2 Optimal Target Symbol Scaling

Derivation of solution to

$$\underset{\mu}{\text{minimise}} \quad \left\|\mathbf{W}^{(\text{ZF})}\mathbf{H}\big(\mathbf{AW}-\mu\mathbf{W}\big)\right\|^2. \tag{12}$$

Since $\mathbf{W}^{(\text{ZF})}\mathbf{H}$ is an orthogonal projection into the channel space, which $\mathbf{AW}$ and $\mathbf{W}$ inherently lie within, this reduces to

$$\underset{\mu}{\text{minimise}} \quad \left\|\mathbf{AW}-\mu\mathbf{W}\right\|^2. \tag{13}$$

This can be written,

$$f(\mu) = \text{Tr}\left((\mathbf{A}-\mu\mathbf{I}_M)\mathbf{W}\mathbf{W}^\dagger(\mathbf{A}^\dagger-\mu^*\mathbf{I}_M)\right). \tag{14}$$

and is minimised by setting [170]

$$\frac{\partial f(\mu)}{\partial \mu^*} = \text{Tr}\left((\mathbf{A}-\mu\mathbf{I}_M)\mathbf{W}\mathbf{W}^\dagger\right) = 0 \tag{15}$$

resulting in

$$\mu = \frac{\text{Tr}\left(\mathbf{W}^\dagger\mathbf{AW}\right)}{\text{Tr}\left(\mathbf{W}^\dagger\mathbf{W}\right)} = \frac{\sum_{k=1}^K \mathbf{w}_k^\dagger\mathbf{A}\mathbf{w}_k}{\sum_{k=1}^K \mathbf{w}_k^\dagger\mathbf{w}_k}. \tag{16}$$

2 Transform Coding-based Signal Compression for Uplink MIMO C-RAN

2.1 Upper Bound on Sample & Forward Capacity

$$\begin{aligned}
\mathcal{C}_{\text{SUM}}^{\text{SF}} &= \log_2\det\left(\mathbf{I}_K+\rho\mathbf{H}^\dagger\big(\mathbf{\Psi}+\mathbf{I}_M\big)^{-1}\mathbf{H}\right) && (17)\\
&< \log_2\det\left(\mathbf{I}_K+\rho\mathbf{H}^\dagger\Big((\rho\mathbf{D}_{\text{H}}+\mathbf{I}_M)2^{-\mathcal{R}/M}+\mathbf{I}_M\Big)^{-1}\mathbf{H}\right) && (18)\\
&< \log_2\det\left(\mathbf{I}_K+2^{\mathcal{R}/M}\rho\mathbf{H}^\dagger\big(\rho\mathbf{D}_{\text{H}}+\mathbf{I}_M\big)^{-1}\mathbf{H}\right) && (19)\\
&< \lim_{\rho\to\infty}\log_2\det\left(\mathbf{I}_K+2^{\mathcal{R}/M}\rho\mathbf{H}^\dagger\big(\rho\mathbf{D}_{\text{H}}+\mathbf{I}_M\big)^{-1}\mathbf{H}\right) && (20)\\
&= \log_2\det\left(\mathbf{I}_K+2^{\mathcal{R}/M}\mathbf{H}^\dagger\mathbf{D}_{\text{H}}^{-1}\mathbf{H}\right) && (21)\\
&\leq \frac{\mathcal{R}K}{M}+\log_2\det\left(\mathbf{H}^\dagger\mathbf{D}_{\text{H}}^{-1}\mathbf{H}\right)+2^{-\mathcal{R}/M}\log_2(e)\,\text{Tr}\left(\left(\mathbf{H}^\dagger\mathbf{D}_{\text{H}}^{-1}\mathbf{H}\right)^{-1}\right), && (22)
\end{aligned}$$

where (18) follows from

$$\mathbf{\Psi} = \big(\rho\mathbf{D}_{\text{H}}+\mathbf{I}_M\big)\frac{1}{2^{\mathcal{R}/M}-1} > \big(\rho\mathbf{D}_{\text{H}}+\mathbf{I}_M\big)2^{-\mathcal{R}/M}, \tag{23}$$

and (22) from the identity

$$\log_2\det\left(\mathbf{I}_K+\mathbf{A}\right) \leq \log_2\det\left(\mathbf{A}\right)+\log_2(e)\,\text{Tr}\left(\mathbf{A}^{-1}\right). \tag{24}$$

2.2 SCA - Joint Rate Allocation

Aiming to solve

$$\begin{aligned} \underset{r_i}{\text{minimise}} \quad & \sum_{i=1}^{t} \frac{q_i}{2^{r_i}-1} \\ \text{subject to} \quad & \sum_{i=1}^{t} r_i \leq \mathcal{R}. \end{aligned} \tag{25}$$

To solve, easier to make substitution,

$$2^{r_i} = e^{x_i} \tag{26}$$

i.e $x_i = r_i \ln(2)$, and

$$\sum_{i=1}^{t} x_i = \mathcal{R}\ln(2). \tag{27}$$

The Lagrangian is then

$$\mathcal{L} = \sum_{i=1}^{t} \frac{q_i}{e^{x_i}-1} + \zeta\big(\sum_{i=1}^{t} x_i - \mathcal{R}\ln(2)\big) \tag{28}$$

giving

$$\frac{\partial \mathcal{L}}{\partial x_i} = \frac{q_i}{2-2\cosh(x_i)} + \zeta = 0 \tag{29}$$

Or

$$\cosh(x_i) = 1 + \frac{q_i}{2\zeta} \tag{30}$$

leading to

$$x_i = \ln\left(2\zeta + q_i + \sqrt{q_i(q_i+4\zeta)}\right) - \ln(2\zeta) \tag{31}$$

Resubstituting for r_i and setting $\gamma = 2\zeta$,

$$r_i = \left[\log_2\left(\gamma + q_i + \sqrt{q_i}\sqrt{q_i+2\gamma}\right) - \log_2(\gamma)\right]^+ \tag{32}$$

where $\gamma \in \mathbb{R}^+$ is chosen such that the fronthaul constraint is met.

3 Dimension Reduction for Distributed MIMO C-RAN

3.1 Conditional Entropy

This can be shown by deriving the conditional distribution, but easiest to illustrate by working backwards

$$\mathcal{I}(\mathbf{z}_l; \mathbf{x}|\mathbf{z}_l^{\mathsf{c}}) = \mathcal{H}\big(\mathbf{z}_l|\mathbf{z}_l^{\mathsf{c}}\big) - \mathcal{H}\big(\mathbf{z}_l|\mathbf{x}, \mathbf{z}_l^{\mathsf{c}}\big) \tag{33}$$

$$= \mathcal{H}\big(\mathbf{z}_l|\mathbf{z}_l^{\mathsf{c}}\big) - \mathcal{H}\big(\boldsymbol{\eta}_{\text{A}}\big) \tag{34}$$

$$= \mathcal{H}\big(\mathbf{z}_l|\mathbf{z}_l^{\mathsf{c}}\big) - \log_2(2\pi e)^N \tag{35}$$

From standard information theory properties

$$\mathcal{I}(\mathbf{z}_1, \ldots, \mathbf{z}_L; \mathbf{x}) = \mathcal{I}(\mathbf{z}_l^{\mathsf{c}}; \mathbf{x}) + \mathcal{I}(\mathbf{z}_l; \mathbf{x} | \mathbf{z}_l^{\mathsf{c}}). \tag{36}$$

Applying the matrix determinant lemma [170], $\det(\mathbf{A} + \mathbf{BC}) = \det(\mathbf{I} + \mathbf{CA}^{-1}\mathbf{B}) \det(\mathbf{A})$,

$$\begin{aligned} \mathcal{I}(\mathbf{z}_1, \ldots, \mathbf{z}_L; \mathbf{x}) &= \log_2 \det \left(\mathbf{I}_K + \rho \sum_{i=1}^{l} \mathbf{H}_i^{\dagger} \mathbf{A}_i \mathbf{A}_i^{\dagger} \mathbf{H}_i \right) \\ &= \log_2 \det \left(\mathbf{I}_N + \rho \mathbf{A}_l^{\dagger} \mathbf{H}_l \mathbf{Q}_l \mathbf{H}_l^{\dagger} \mathbf{A}_l \right) \\ &\quad + \log_2 \det \left(\mathbf{Q}_l^{-1} \right) \end{aligned} \tag{37}$$

where

$$\mathbf{Q}_l = \left(\mathbf{I}_K + \sum_{i \neq l} \mathbf{H}_i^{\dagger} \mathbf{A}_i \mathbf{A}_i^{\dagger} \mathbf{H}_i \right)^{-1}. \tag{38}$$

By inspection of (36) and (37)

$$\mathcal{I}(\mathbf{z}_l; \mathbf{x} | \mathbf{z}_l^{\mathsf{c}}) = \log_2 \det \left(\mathbf{I}_N + \rho \mathbf{A}_l^{\dagger} \mathbf{H}_l \mathbf{Q}_l \mathbf{H}_l^{\dagger} \mathbf{A}_l \right), \tag{39}$$

and

$$\mathcal{H}\left(\mathbf{z}_l | \mathbf{z}_l^{\mathsf{c}} \right) = \log_2 \det \left(\mathbf{I}_N + \rho \mathbf{A}_l^{\dagger} \mathbf{H}_l \mathbf{Q}_l \mathbf{H}_l^{\dagger} \mathbf{A}_l \right) + \log_2 (2\pi e)^N. \tag{40}$$

3.2 Determinant Maximisation

Consider the matrix product

$$\mathbf{A}^{\dagger} \mathbf{B} \mathbf{A} \tag{41}$$

where $\mathbf{B} \in \mathbb{C}^{n \times n}$ is a Hermitian symmetric matrix, and $\mathbf{A} \in \mathbb{C}^{m \times n}$, $m \leq n$, a rectangular matrix with orthonormal columns,

$$\mathbf{A}^{\dagger} \mathbf{A} = \mathbf{I}_m. \tag{42}$$

By the Poincaré separation theorem [14], the eigenvalues of $\mathbf{A}^{\dagger} \mathbf{B} \mathbf{A}$ can be upper bounded

$$\alpha_i \leq \beta_i \tag{43}$$

where α_i and β_i are the ordered eigenvalues of $\mathbf{A}^{\dagger} \mathbf{B} \mathbf{A}$ and $\mathbf{B}$, respectively. We therefore have

$$\det \left(\mathbf{A}^{\dagger} \mathbf{B} \mathbf{A} \right) = \prod_{i=1}^{m} \alpha_i \leq \prod_{i=1}^{m} \beta_i. \tag{44}$$

Setting the columns of $\mathbf{A}$ to be the m principal eigenvectors of $\mathbf{B}$ achieves equality in (44). This $\mathbf{A}$ is non-unique, since any

$$\mathbf{A}^{\star} = \mathbf{A} \boldsymbol{\Theta} \tag{45}$$

where $\boldsymbol{\Theta} \in \mathbb{C}^{n \times n}$ is a unitary matrix also achieves equality with the upper bound.

Bibliography

[1] Andrea Abrardo, Gabor Fodor, Marco Moretti, and Miklos Telek. MMSE Receiver Design and SINR Calculation in MU-MIMO Systems With Imperfect CSI. IEEE Wireless Communications Letters, 8(1):269–272, 2019.

[2] M. Agiwal, A. Roy, and N. Saxena. Next Generation 5G Wireless Networks: A Comprehensive Survey. IEEE Communications Surveys Tutorials, 18(3):1617–1655, 2016.

[3] S. M. Alamouti. A Simple Transmit Diversity Technique for Wireless Communications. IEEE Journal on Selected Areas in Communications, 16(8):1451–1458, 1998.

[4] P. V. Amadori and C. Masouros. Constant Envelope Precoding by Interference Exploitation in Phase Shift Keying-Modulated Multiuser Transmission. IEEE Transactions on Wireless Communications, 16(1):538–550, 2017.

[5] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang. What Will 5G Be? IEEE Journal on Selected Areas in Communications, 32(6):1065–1082, 2014.

[6] J. Armstrong. Peak-to-Average Power Reduction for OFDM by Repeated Clipping and Frequency Domain Filtering. Electronics Letters, 38(5):246–247, Feb 2002.

[7] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske. How Much Energy Is Needed to Run a Wireless Network? IEEE Wireless Communications, 18(5):40–49, 2011.

[8] K. Bae, J. G. Andrews, and E. J. Powers. Adaptive Active Constellation Extension Algorithm for Peak-to-Average Ratio Reduction in OFDM. IEEE Communications Letters, 14(1):39–41, 2010.

[9] H. Bao, J. Fang, Z. Chen, H. Li, and S. Li. An Efficient Bayesian PAPR Reduction Method for OFDM-Based Massive MIMO Systems. IEEE Transactions on Wireless Communications, 15(6):4183–4195, 2016.

[10] Dror Baron, Marco F. Duarte, Michael B. Wakin, Shriram Sarvotham, and Richard G. Baraniuk. Distributed Compressive Sensing, 2009.

[11] M. Bashar, K. Cumanan, A. G. Burr, M. Debbah, and H. Q. Ngo. Enhanced Max-Min SINR for Uplink Cell-Free Massive MIMO Systems. In 2018 IEEE International Conference on Communications (ICC), pages 1–6, 2018.

[12] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, and M. Debbah. Cell-Free Massive MIMO With Limited Backhaul. In 2018 IEEE International Conference on Communications (ICC), pages 1–7, 2018.

[13] R. W. Bauml, R. F. H. Fischer, and J. B. Huber. Reducing the Peak-to-Average Power Ratio of Multicarrier Modulation by Selected Mapping. Electronics Letters, 32(22):2056–2057, 1996.

[14] Richard Bellman. Introduction to Matrix Analysis. SIAM, 1997.

[15] Emil Björnson, Jakob Hoydis, Luca Sanguinetti, and Others. Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency. Now Publishers, Inc., 2017.

[16] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah. Massive MIMO Systems With Non-Ideal Hardware: Energy Efficiency, Estimation, and Capacity Limits. IEEE Transactions on Information Theory, 60(11):7112–7139, 2014.

[17] E. Björnson, J. Hoydis, and L. Sanguinetti. Massive MIMO Has Unlimited Capacity. IEEE Transactions on Wireless Communications, 17(1):574–590, 2018.

[18] E. Björnson, E. G. Larsson, and M. Debbah. Massive MIMO for Maximal Spectral Efficiency: How Many Users and Pilots Should Be Allocated? IEEE Transactions on Wireless Communications, 15(2):1293–1308, 2016.

[19] E. Björnson, E. G. Larsson, and T. L. Marzetta. Massive MIMO: Ten Myths and One Critical Question. IEEE Communications Magazine, 54(2):114–123, 2016.

[20] E. Björnson and L. Sanguinetti. Making Cell-Free Massive MIMO Competitive With MMSE Processing and Centralized Implementation. IEEE Transactions on Wireless Communications, 19(1):77–90, 2020.

[21] E. Björnson, R. Zakhour, D. Gesbert, and B. Ottersten. Cooperative Multicell Precoding: Rate Region Characterization and Distributed Strategies With Instantaneous and Statistical CSI. IEEE Transactions on Signal Processing, 58(8):4298–4310, 2010.

[22] F. Boccardi, B. Clerckx, A. Ghosh, E. Hardouin, G. Jöngren, K. Kusume, E. Onggosanusi, and Y. Tang. Multiple-Antenna Techniques in LTE-Advanced. IEEE Communications Magazine, 50(3):114–121, 2012.

[23] R. Brandt and M. Bengtsson. Distributed CSI Acquisition and Coordinated Precoding for TDD Multicell MIMO Systems. IEEE Transactions on Vehicular Technology, 65(5):2890–2906, 2016.

[24] D. G. Brennan. Linear Diversity Combining Techniques. Proceedings of the IRE, 47(6):1075–1102, 1959.

[25] K. Bumman, M. Junghwan, and K. Ildu. Efficiently Amplified. IEEE Microwave Magazine, 11(5):87–100, 2010.

[26] Julian Jakob Bussgang. Crosscorrelation Functions of Amplitude-Distorted Gaussian Signals. 1952.

[27] E. J. Candes, J. Romberg, and T. Tao. Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information. IEEE Transactions on Information Theory, 52(2):489–509, 2006.

[28] E. J. Candes and T. Tao. Decoding by Linear Programming. IEEE Transactions on Information Theory, 51(12):4203–4215, 2005.

[29] E. J. Candes and M. B. Wakin. An Introduction to Compressive Sampling. IEEE Signal Processing Magazine, 25(2):21–30, 2008.

[30] A. Cavalcante, I. Almeida, E. Medeiros, and M. Berg. A Novel Model of Beamspace Beamforming Coefficients for Fronthaul Load Reduction. In ICC 2019 - 2019 IEEE International Conference on Communications (ICC), pages 1–6, 2019.

[31] H. Cha, H. Chae, K. Kim, J. Jang, J. Yang, and D. K. Kim. Generalized Inverse Aided PAPR-Aware Linear Precoder Design for MIMO-OFDM System. IEEE Communications Letters, 18(8):1363–1366, 2014.

[32] W. Chang, T. Xie, F. Zhou, J. Tian, and X. Zhang. A Prefiltering C-RAN Architecture With Compressed Link Data Rate in Massive MIMO. In 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring), pages 1–6, 2016.

[33] J. Chen, C. Wang, K. Wong, and C. Wen. Low-Complexity Precoding Design for Massive Multiuser MIMO Systems Using Approximate Message Passing. IEEE Transactions on Vehicular Technology, 65(7):5707–5714, 2016.

[34] J. Chen, C. Wen, and K. Wong. Improved Constant Envelope Multiuser Precoding for Massive MIMO Systems. IEEE Communications Letters, 18(8):1311–1314, Aug 2014.

[35] Z. Chen and E. Björnson. Channel Hardening and Favorable Propagation in Cell-Free Massive MIMO With Stochastic Geometry. IEEE Transactions on Communications, 66(11):5205–5219, 2018.

[36] Cheong Yui Wong, R. S. Cheng, K. B. Lataief, and R. D. Murch. Multiuser OFDM With Adaptive Subcarrier, Bit, and Power Allocation. IEEE Journal on Selected Areas in Communications, 17(10):1747–1758, 1999.

[37] Mung Chiang, Prashanth Hande, Tian Lan, Chee Wei Tan, and Others. Power Control in Wireless Cellular Networks. Foundations and Trends in Networking, 2(4):381–533, 2008.

[38] D. Chizhik, J. Ling, P. W. Wolniansky, R. A. Valenzuela, N. Costa, and K. Huber. Multiple-Input-Multiple-Output Measurements and Modeling in Manhattan. IEEE Journal on Selected Areas in Communications, 21(3):321–331, 2003.

[39] J. Choi, B. L. Evans, and A. Gatherer. Space-Time Fronthaul Compression of Complex Baseband Uplink LTE Signals. In 2016 IEEE International Conference on Communications (ICC), pages 1–6, 2016.

[40] J. Choi, D. J. Love, and P. Bidigare. Downlink Training Techniques for FDD Massive MIMO Systems: Open-Loop and Closed-Loop Training With Memory. IEEE Journal of Selected Topics in Signal Processing, 8(5):802–814, 2014.

[41] L. Cimini. Analysis and Simulation of a Digital Mobile Channel Using Orthogonal Frequency Division Multiplexing. IEEE Transactions on Communications, 33(7):665–675, 1985.

[42] L. J. Cimini and N. R. Sollenberger. Peak-to-Average Power Ratio Reduction of an OFDM Signal Using Partial Transmit Sequences. IEEE Communications Letters, 4(3):86–88, 2000.

[43] Bruno Clerckx and Claude Oestges. MIMO Wireless Networks: Channels, Techniques and Standards for Academic Press, 2013.

[44] CPRI Consortium and Others. eCPRI Specification V1. 0, 2017.

[45] M. Costa. Writing on Dirty Paper (Corresp.). IEEE Transactions on Information Theory, 29(3):439–441, 1983.

[46] Romain Couillet and Merouane Debbah. Random Matrix Methods for Wireless Communications. Cambridge University Press, 2011.

[47] Thomas M Cover and Joy A Thomas. Elements of Information Theory. John Wiley & Sons, 2012.

[48] B. Dai and W. Yu. Sparse Beamforming and User-Centric Clustering for Downlink Cloud Radio Access Network. IEEE Access, 2:1326–1339, 2014.

[49] A. Del Coso and S. Simoens. Distributed Compression for MIMO Coordinated Networks With a Backhaul Constraint. IEEE Transactions on Wireless Communications, 8(9):4698–4709, 2009.

[50] O. T. Demir and E. Bjornson. The Bussgang Decomposition of Nonlinear Systems: Basic Theory and MIMO Extensions [Lecture Notes]. IEEE Signal Processing Magazine, 38(1):131–136, 2021.

[51] M. F. Duarte, S. Sarvotham, D. Baron, M. B. Wakin, and R. G. Baraniuk. Distributed Compressed Sensing of Jointly Sparse Signals. In Conference Record of the Thirty-Ninth Asilomar Conference onSignals, Systems and Computers, 2005., pages 1537–1541, 2005.

[52] U. Dötsch, M. Doll, H. Mayer, F. Schaich, J. Segel, and P. Sehier. Quantitative Analysis of Split Base Station Processing and Determination of Advantageous Architectures for LTE. Bell Labs Technical Journal, 18(1):105–128, 2013.

[53] Abbas El Gamal and Young-Han Kim. Network Information Theory. Cambridge University Press, 2011.

[54] O. Elijah, C. Y. Leow, T. A. Rahman, S. Nunoo, and S. Z. Iliya. A Comprehensive Survey of Pilot Contamination in Massive MIMO—5G System. IEEE Communications Surveys Tutorials, 18(2):905–923, 2016.

[55] D. Falconer, S. L. Ariyavisitakul, A. Benyamin-Seeyar, and B. Eidson. Frequency Domain Equalization for Single-Carrier Broadband Wireless Systems. IEEE Communications Magazine, 40(4):58–66, 2002.

[56] Imola K Fodor. A Survey of Dimension Reduction Techniques. Technical report, Lawrence Livermore National Lab., CA (US), 2002.

[57] Global Mobile Data Traffic Forecast. Cisco Visual Networking Index: Forecast and Methodology, 2008–2013. Update, 14, 2009.

[58] Global Mobile Data Traffic Forecast. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022. Update, 2017, 2019.

[59] P Frenger, J Hederen, M Hessler, and G Interdonato. Improved Antenna Arrangement for Distributed Massive MIMO. WO Patent Application, 2018103897, 2017. `https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2018103897`.

[60] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson. Linear Pre-Coding Performance in Measured Very-Large MIMO Channels. In 2011 IEEE Vehicular Technology Conference (VTC Fall), pages 1–5, 2011.

[61] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson. Massive MIMO Performance Evaluation Based on Measured Propagation Data. IEEE Transactions on Wireless Communications, 14(7):3899–3911, 2015.

[62] M. Gastpar, P. L. Dragotti, and M. Vetterli. The Distributed Karhunen–Loève Transform. IEEE Transactions on Information Theory, 52(12):5177–5196, 2006.

[63] D. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, O. Simeone, and W. Yu. Multi-Cell MIMO Cooperative Networks: A New Look at Interference. IEEE Journal on Selected Areas in Communications, 28(9):1380–1408, 2010.

[64] D. Gesbert, M. Kountouris, R. W. Heath, C. Chae, and T. Salzer. Shifting the MIMO Paradigm. IEEE Signal Processing Magazine, 24(5):36–46, Sep. 2007.

[65] M. Gharavi-Alkhansari and A. B. Gershman. Fast Antenna Subset Selection in MIMO Systems. IEEE Transactions on Signal Processing, 52(2):339–347, 2004.

[66] A. Ghosh, A. Maeder, M. Baker, and D. Chandramouli. 5G Evolution: A View on 5G Cellular Technology Beyond 3GPP Release 15. IEEE Access, 7:127639–127651, 2019.

[67] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, and C. E. Wheatley. On the Capacity of a Cellular CDMA System. IEEE Transactions on Vehicular Technology, 40(2):303–312, 1991.

[68] G. Ginis and J. M. Cioffi. On the Relation Between V-BLAST and the GDFE. IEEE Communications Letters, 5(9):364–366, 2001.

[69] H. Gish and J. Pierce. Asymptotically Efficient Quantizing. IEEE Transactions on Information Theory, 14(5):676–683, 1968.

[70] A. J. Goldsmith and P. P. Varaiya. Capacity of Fading Channels With Channel Side Information. IEEE Transactions on Information Theory, 43(6):1986–1992, 1997.

[71] A. Gorokhov, D. A. Gore, and A. J. Paulraj. Receive Antenna Selection for MIMO Spatial Multiplexing: Theory and Algorithms. IEEE Transactions on Signal Processing, 51(11):2796–2807, 2003.

[72] V. K. Goyal. Theoretical Foundations of Transform Coding. IEEE Signal Processing Magazine, 18(5):9–21, 2001.

[73] R. Gray. Vector Quantization. IEEE ASSP Magazine, 1(2):4–29, 1984.

[74] R. Gray and L. Davisson. Quantizer Mismatch. IEEE Transactions on Communications, 23(4):439–443, 1975.

[75] Charles Miller Grinstead and James Laurie Snell. Introduction to Probability. American Mathematical Soc., 2012.

[76] GSMA. Improving Wireless Connectivity Through Small Cell Deployment. GSMA White Paper, 2016.

[77] J. Hamkins and K. Zeger. Gaussian Source Coding With Spherical Codes. IEEE Transactions on Information Theory, 48(11):2980–2989, 2002.

[78] Hangjun Chen and A. M. Haimovich. Iterative Estimation and Cancellation of Clipping Noise for OFDM Signals. IEEE Communications Letters, 7(7):305–307, 2003.

[79] P. Harris, W. B. Hasan, S. Malkowsky, J. Vieira, S. Zhang, M. Beach, L. Liu, E. Mellios, A. Nix, S. Armour, A. Doufexi, K. Nieman, and N. Kundargi. Serving 22 Users in Real-Time With a 128-Antenna Massive MIMO Testbed. In 2016 IEEE International Workshop on Signal Processing Systems (SiPS), pages 266–272, 2016.

[80] P. Harris, S. Malkowsky, J. Vieira, E. Bengtsson, F. Tufvesson, W. B. Hasan, L. Liu, M. Beach, S. Armour, and O. Edfors. Performance Characterization of a Real-Time Massive MIMO System With LOS Mobile Channels. IEEE Journal on Selected Areas in Communications, 35(6):1244–1253, 2017.

[81] P. Harris, S. Zang, A. Nix, M. Beach, S. Armour, and A. Doufexi. A Distributed Massive MIMO Testbed to Assess Real-World Performance and Feasibility. In 2015 IEEE 81st Vehicular Technology Conference (VTC Spring), pages 1–2, 2015.

[82] P. Harris, S. Zhang, M. Beach, E. Mellios, A. Nix, S. Armour, A. Doufexi, K. Nieman, and N. Kundargi. LOS Throughput Measurements in Real-Time With a 128-Antenna Massive MIMO Testbed. In 2016 IEEE Global Communications Conference (GLOBECOM), pages 1–7, 2016.

[83] Z. Hasan, H. Boostanimehr, and V. K. Bhargava. Green Cellular Networks: A Survey, Some Research Issues and Challenges. IEEE Communications Surveys Tutorials, 13(4):524–540, 2011.

[84] B. Hassibi and B. M. Hochwald. How Much Training Is Needed in Multiple-Antenna Wireless Links? IEEE Transactions on Information Theory, 49(4):951–963, April 2003.

[85] B. Hassibi and H. Vikalo. On the Sphere-Decoding Algorithm I. Expected Complexity. IEEE Transactions on Signal Processing, 53(8):2806–2818, Aug 2005.

[86] R. Heath, S. Peters, Y. Wang, and J. Zhang. A Current Perspective on Distributed Antenna Systems for the Downlink of Cellular Systems. IEEE Communications Magazine, 51(4):161–167, 2013.

[87] R. W. Heath, S. Sandhu, and A. Paulraj. Antenna Selection for Spatial Multiplexing Systems With Linear Receivers. IEEE Communications Letters, 5(4):142–144, April 2001.

[88] A. Hedayat and A. Nosratinia. Outage and Diversity of Linear Receivers in Flat-Fading MIMO Channels. IEEE Transactions on Signal Processing, 55(12):5868–5873, 2007.

[89] B. M. Hochwald, C. B. Peel, and A. L. Swindlehurst. A Vector-Perturbation Technique for Near-Capacity Multiantenna Multiuser Communication-Part II: Perturbation. IEEE Transactions on Communications, 53(3):537–544, 2005.

[90] M. Hong, R. Sun, H. Baligh, and Z. Luo. Joint Base Station Clustering and Beamformer Design for Partial Coordinated Transmission in Heterogeneous Networks. IEEE Journal on Selected Areas in Communications, 31(2):226–240, 2013.

[91] Roger a Horn and Charles R Johnson. Matrix Analysis. Cambridge University Press, 2012.

[92] Harold Hotelling. Analysis of a Complex of Statistical Variables Into Principal components. Journal of Educational Psychology, 24(6):417, 1933.

[93] J. Hoydis, C. Hoek, T. Wild, and S. Ten Brink. Channel Measurements for Large Antenna Arrays. In 2012 International Symposium on Wireless Communication Systems (ISWCS), pages 811–815, 2012.

[94] J. Hoydis, S. Ten Brink, and M. Debbah. Massive MIMO in the UL/DL of Cellular Networks: How Many Antennas Do We Need? IEEE Journal on Selected Areas in Communications, 31(2):160–171, 2013.

[95] Y. Huang, C. Lu, M. Berg, and P. Ödling. Functional Split of Zero-Forcing Based Massive MIMO for Fronthaul Load Reduction. IEEE Access, 6:6350–6359, 2018.

[96] H. Huh, A. M. Tulino, and G. Caire. Network MIMO With Linear Zero-Forcing Beamforming: Large System Analysis, Impact of Channel Estimation, and Reduced-Complexity Scheduling. IEEE Transactions on Information Theory, 58(5):2911–2934, 2012.

[97] T. Hwang, C. Yang, G. Wu, S. Li, and G. Ye Li. OFDM and Its Wireless Applications: A Survey. IEEE Transactions on Vehicular Technology, 58(4):1673–1694, 2009.

[98] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov. 5G Backhaul Challenges and Emerging Research Directions: A Survey. IEEE Access, 4:1743–1766, 2016.

[99] S. Jacobsson, Y. Ettefagh, G. Durisi, and C. Studer. All-Digital Massive Mimo With a Fronthaul Constraint. In 2018 IEEE Statistical Signal Processing Workshop (SSP), pages 218–222, 2018.

[100] J. Jelitto and G. Fettweis. Reduced Dimension Space-Time Processing for Multi-Antenna Wireless Systems. IEEE Wireless Communications, 9(6):18–25, 2002.

[101] T. Jiang and Y. Wu. An Overview: Peak-to-Average Power Ratio Reduction Techniques for OFDM Signals. IEEE Transactions on Broadcasting, 54(2):257–268, 2008.

[102] M. Joham, W. Utschick, and J. A. Nossek. Linear Transmit Processing in MIMO Communications Systems. IEEE Transactions on Signal Processing, 53(8):2700–2712, 2005.

[103] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath. Pilot Contamination and Precoding in Multi-Cell TDD Systems. IEEE Transactions on Wireless Communications, 10(8):2640–2651, 2011.

[104] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfröjd, and T. Svensson. The Role of Small Cells, Coordinated Multipoint, and Massive MIMO in 5G. IEEE Communications Magazine, 52(5):44–51, 2014.

[105] T. Kageyama, O. Muta, C. Chen, and S. Pollin. Effect of Limiter Based PAPR Reduction for Massive MIMO Systems. In 2018 International Japan-Africa Conference on Electronics, Communications and Computations (JAC-ECC), pages 43–46, 2018.

[106] F. Kaltenberger, D. Gesbert, R. Knopp, and M. Kountouris. Correlation and Capacity of Measured Multi-User MIMO Channels. In 2008 IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications, pages 1–5, 2008.

[107] K. Kang, Z. Fang, H. Wang, H. Qian, and Y. Yang. Dummy Signal Precoding for PAPR Reduction in MIMO Communication System. In 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP), pages 1–5, 2018.

[108] K. Karakayali, R. Yates, G. Foschini, and R. Valenzuela. Optimum Zero-Forcing Beamforming With Per-Antenna Power Constraints. In 2007 IEEE International Symposium on Information Theory, pages 101–105, 2007.

[109] S. Khademi, A. Van Der Veen, and T. Svantesson. Precoding Technique for Peak-to-Average-Power-Ratio (PAPR) Reduction in MIMO OFDM/A Systems. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3005–3008, 2012.

[110] J. Kim, S. Park, O. Simeone, I. Lee, and S. Shamai. Joint Design of Digital and Analog Processing for Downlink C-RAN With Large-Scale Antenna Arrays. In 2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pages 1–5, 2017.

[111] K. Kim, S. Myung, S. Park, J. Lee, M. Kan, Y. Shinohara, J. Shin, and J. Kim. Low-Density Parity-Check Codes for ATSC 3.0. IEEE Transactions on Broadcasting, 62(1):189–196, 2016.

[112] B. S. Krongold and D. L. Jones. PAR Reduction in OFDM via Active Constellation Extension. IEEE Transactions on Broadcasting, 49(3):258–268, 2003.

[113] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta. Massive MIMO for Next Generation Wireless Systems. IEEE Communications Magazine, 52(2):186–195, 2014.

[114] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, and K. Sayana. Coordinated Multipoint Transmission and Reception in LTE-Advanced: Deployment Scenarios and Operational Challenges. IEEE Communications Magazine, 50(2):148–155, 2012.

[115] H. Li, L. Han, R. Duan, and G. M. Garner. Analysis of the Synchronization Requirements of 5g and Corresponding Solutions. IEEE Communications Standards Magazine, 1(1):52–58, 2017.

[116] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu. Channel Estimation and Performance Analysis of One-Bit Massive MIMO Systems. IEEE Transactions on Signal Processing, 65(15):4075–4089, 2017.

[117] C. Lim, T. Yoo, B. Clerckx, B. Lee, and B. Shim. Recent Trend of Multiuser MIMO in LTE-Advanced. IEEE Communications Magazine, 51(3):127–135, 2013.

[118] A. Liu, X. Chen, W. Yu, V. K. N. Lau, and M. Zhao. Two-Timescale Hybrid Compression and Forward for Massive MIMO Aided C-RAN. IEEE Transactions on Signal Processing, 67(9):2484–2498, 2019.

[119] A. Liu and V. K. N. Lau. Two-Stage Constant-Envelope Precoding for Low-Cost Massive MIMO Systems. IEEE Transactions on Signal Processing, 64(2):485–494, 2016.

[120] J. Liu, A. Liu, and V. K. N. Lau. Compressive Interference Mitigation and Data Recovery in Cloud Radio Access Networks With Limited Fronthaul. IEEE Transactions on Signal Processing, 65(6):1437–1446, 2017.

[121] L. Liu, S. Bi, and R. Zhang. Joint Power Control and Fronthaul Rate Allocation for Throughput Maximization in OFDMA-Based Cloud Radio Access Network. IEEE Transactions on Communications, 63(11):4097–4110, 2015.

[122] L. Liu, R. Chen, S. Geirhofer, K. Sayana, Z. Shi, and Y. Zhou. Downlink MIMO in LTE-Advanced: SU-MIMO vs. MU-MIMO. IEEE Communications Magazine, 50(2):140–147, 2012.

[123] L. Liu and R. Zhang. Optimized Uplink Transmission in Multi-Antenna C-RAN With Spatial Compression and Forward. IEEE Transactions on Signal Processing, 63(19):5083–5095, 2015.

[124] Lizhong Zheng and D. N. C. Tse. Diversity and Multiplexing: A Fundamental Tradeoff in Multiple-Antenna Channels. IEEE Transactions on Information Theory, 49(5):1073–1096, 2003.

[125] S. Loyka and F. Gagnon. Performance Analysis of the V-BLAST Algorithm: An Analytical Approach. IEEE Transactions on Wireless Communications, 3(4):1326–1337, 2004.

[126] S. Luo, R. Zhang, and T. J. Lim. Downlink and Uplink Energy Minimization Through User Association and Beamforming in C-RAN. IEEE Transactions on Wireless Communications, 14(1):494–508, 2015.

[127] Luqing Wang and C. Tellambura. A Simplified Clipping and Filtering Technique for PAR Reduction in OFDM Systems. IEEE Signal Processing Letters, 12(6):453–456, 2005.

[128] H. Ma, B. Wang, and K. J. R. Liu. Distributed Signal Compressive Quantization and Parallel Interference Cancellation for Cloud Radio Access Network. IEEE Transactions on Communications, 66(9):4186–4198, 2018.

[129] A. Maeder, M. Lalam, A. De Domenico, E. Pateromichelakis, D. Wübben, J. Bartelt, R. Fritzsche, and P. Rost. Towards a Flexible Functional Split for Cloud-RAN Networks. In 2014 European Conference on Networks and Communications (EuCNC), pages 1–5, 2014.

[130] S. Malkowsky, J. Vieira, L. Liu, P. Harris, K. Nieman, N. Kundargi, I. C. Wong, F. Tufvesson, V. Öwall, and O. Edfors. The World's First Real-Time Testbed for Massive MIMO: Design, Implementation, and Validation. IEEE Access, 5:9073–9088, 2017.

[131] M. W. Marcellin and T. R. Fischer. Trellis Coded Quantization of Memoryless and Gauss-Markov Sources. IEEE Transactions on Communications, 38(1):82–93, 1990.

[132] D. Maryopi, M. Bashar, and A. Burr. On the Uplink Throughput of Zero Forcing in Cell-Free Massive MIMO With Coarse Quantization. IEEE Transactions on Vehicular Technology, 68(7):7220–7224, 2019.

[133] T. L. Marzetta. Noncooperative Cellular Wireless With Unlimited Numbers of Base Station Antennas. IEEE Transactions on Wireless Communications, 9(11):3590–3600, 2010.

[134] T. L. Marzetta and B. M. Hochwald. Fast Transfer of Channel State Information in Wireless Systems. IEEE Transactions on Signal Processing, 54(4):1268–1278, 2006.

[135] Thomas L Marzetta. Fundamentals of Massive MIMO. Cambridge University Press, 2016.

[136] Y. Matsumoto, K. Tateishi, and K. Higuchi. Performance Evaluations on Adaptive PAPR Reduction Method Using Null Space in MIMO Channel for Eigenmode Massive MIMO-OFDM Signals. In 2017 23rd Asia-Pacific Conference on Communications (APCC), pages 1–6, 2017.

[137] Hannah Ritchie Max Roser and Esteban Ortiz-Ospina. Internet. Our World in Data, 2020. https://ourworldindata.org/internet.

[138] A. H. Mehana and A. Nosratinia. Diversity of MMSE MIMO Receivers. IEEE Transactions on Information Theory, 58(11):6788–6805, 2012.

[139] Thomas Melzer. SVD and Its Application to Generalized Eigenvalue Problems. Vienna University of Technology, 2004.

[140] A. Mezghani, M. Rouatbi, and J. A. Nossek. An Iterative Receiver for Quantized MIMO Systems. In 2012 16th IEEE Mediterranean Electrotechnical Conference, pages 1049–1052, 2012.

[141] S. K. Mohammed and E. G. Larsson. Constant-Envelope Multi-User Precoding for Frequency-Selective Massive MIMO Systems. IEEE Wireless Communications Letters, 2(5):547–550, 2013.

[142] S. K. Mohammed and E. G. Larsson. Per-Antenna Constant Envelope Precoding for Large Multi-User MIMO Systems. IEEE Transactions on Communications, 61(3):1059–1071, 2013.

[143] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda. Hybrid Beamforming for Massive MIMO: A Survey. IEEE Communications Magazine, 55(9):134–141, 2017.

[144] Andreas F Molisch. Wireless Communications, volume 34. John Wiley & Sons, 2012.

[145] Andreas F Molisch and Fredrik Tufvesson. Propagation Channel Models for Next-Generation Wireless Communications Systems. IEICE Transactions on Communications, 97(10):2022–2034, 2014.

[146] C. Mollén, J. Choi, E. G. Larsson, and R. W. Heath. Uplink Performance of Wideband Massive MIMO With One-Bit ADCs. IEEE Transactions on Wireless Communications, 16(1):87–100, 2017.

[147] C. Mollén, U. Gustavsson, T. Eriksson, and E. G. Larsson. Spatial Characteristics of Distortion Radiated From Antenna Arrays With Transceiver Nonlinearities. IEEE Transactions on Wireless Communications, 17(10):6663–6679, 2018.

[148] C. Mollén and E. G. Larsson. Multiuser MIMO Precoding With Per-Antenna Continuous-Time Constant-Envelope Constraints. In 2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pages 261–265, 2015.

[149] C. Mollén, E. G. Larsson, and T. Eriksson. Waveforms for the Massive MIMO Downlink: Amplifier Efficiency, Distortion, and Performance. IEEE Transactions on Communications, 64(12):5050–5063, 2016.

[150] S. H. Muller and J. B. Huber. OFDM With Reduced Peak-to-Average Power Ratio by Optimum Combination of Partial Transmit Sequences. Electronics Letters, 33(5):368–369, 1997.

[151] K. Murota and K. Hirade. GMSK Modulation for Digital Mobile Radio Telephony. IEEE Transactions on Communications, 29(7):1044–1050, 1981.

[152] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta. Cell-Free Massive MIMO: Uniformly Great Service for Everyone. In 2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pages 201–205, 2015.

[153] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta. Cell-Free Massive MIMO Versus Small Cells. IEEE Transactions on Wireless Communications, 16(3):1834–1850, 2017.

[154] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta. Energy and Spectral Efficiency of Very Large Multiuser MIMO Systems. IEEE Transactions on Communications, 61(4):1436–1449, April 2013.

[155] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta. Aspects of Favorable Propagation in Massive MIMO. In 2014 22nd European Signal Processing Conference (EUSIPCO), pages 76–80, 2014.

[156] C. Ni, Y. Ma, and T. Jiang. A Novel Adaptive Tone Reservation Scheme for PAPR Reduction in Large-Scale Multi-User MIMO-OFDM Systems. IEEE Wireless Communications Letters, 5(5):480–483, 2016.

[157] K. Niu, K. Chen, J. Lin, and Q. T. Zhang. Polar Codes: Primary Concepts and Practical Decoding Algorithms. IEEE Communications Magazine, 52(7):192–203, 2014.

[158] H. Ochiai and H. Imai. Performance Analysis of Deliberately Clipped OFDM Signals. IEEE Transactions on Communications, 50(1):89–101, Jan 2002.

[159] Ofcom. Communications Market Report. 2018.

[160] Ofcom. Communications Market Report. 2020.

[161] A. P. and R. K. Ganti. QR Approximation for Fronthaul Compression in Uplink Massive MIMO. In 2019 IEEE Globecom Workshops (GC Wkshps), pages 1–7, 2019.

[162] D. P. Palomar and Mung Chiang. A Tutorial on Decomposition Methods for Network Utility Maximization. IEEE Journal on Selected Areas in Communications, 24(8):1439–1451, 2006.

[163] J. Park, S. Park, A. Yazdan, and R. W. Heath. Optimization of Mixed-ADC Multi-Antenna Systems for Cloud-RAN Deployments. IEEE Transactions on Communications, 65(9):3962–3975, 2017.

[164] S. Park, O. Simeone, Y. C. Eldar, and E. Erkip. Optimizing Pilots and Analog Processing for Channel Estimation in Cell-Free Massive MIMO With One-Bit ADCs. In 2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pages 1–5, 2018.

[165] S. Park, O. Simeone, O. Sahin, and S. Shamai. Robust and Efficient Distributed Compression for Cloud Radio Access Networks. IEEE Transactions on Vehicular Technology, 62(2):692–703, 2013.

[166] S. Payami and F. Tufvesson. Channel Measurements and Analysis for Very Large Array Systems at 2.6 GHz. In 2012 6th European Conference on Antennas and Propagation (EUCAP), pages 433–437, 2012.

[167] K. I. Pedersen, P. E. Mogensen, and B. H. Fleury. Power Azimuth Spectrum in Outdoor Environments. Electronics Letters, 33(18):1583–1584, 1997.

[168] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang. Recent Advances in Cloud Radio Access Networks: System Architectures, Key Techniques, and Open Issues. IEEE Communications Surveys Tutorials, 18(3):2282–2308, 2016.

[169] M. Peng, C. Wang, V. Lau, and H. V. Poor. Fronthaul-Constrained Cloud Radio Access Networks: Insights and Challenges. IEEE Wireless Communications, 22(2):152–160, 2015.

[170] Kaare Brandt Petersen and Michael Syskind Pedersen. The Matrix Cookbook, Nov 2012.

[171] Ada SY Poon, Robert W Brodersen, and David NC Tse. Degrees of Freedom in Multiple-Antenna Channels: A Signal Space Approach. IEEE Transactions on Information Theory, 51(2):523–536, 2005.

[172] H. Prabhu, O. Edfors, J. Rodrigues, L. Liu, and F. Rusek. A Low-Complex Peak-to-Average Power Reduction Scheme for OFDM Based Massive MIMO Systems. In 2014 6th International Symposium on Communications, Control and Signal Processing (ISCCSP), pages 114–117, 2014.

[173] J. G. Proakis. Adaptive Equalization for TDMA Digital Mobile Radio. IEEE Transactions on Vehicular Technology, 40(2):333–341, 1991.

[174] John G Proakis and Masoud Salehi. Digital Communications, volume 4. McGraw-Hill New York, 2001.

[175] X. Rao and V. K. N. Lau. Distributed Fronthaul Compression and Joint Signal Recovery in Cloud-RAN. IEEE Transactions on Signal Processing, 63(4):1056–1065, 2015.

[176] F. Rashid-Farrokhi, L. Tassiulas, and K. J. R. Liu. Joint Optimal Power Control and Beamforming in Wireless Networks Using Antenna Arrays. IEEE Transactions on Communications, 46(10):1313–1324, 1998.

[177] T. Richardson and R. Urbanke. The Renaissance of Gallager's Low-Density Parity-Check Codes. IEEE Communications Magazine, 41(8):126–131, 2003.

[178] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson. Scaling Up MIMO: Opportunities and Challenges With Very Large Arrays. IEEE Signal Processing Magazine, 30(1):40–60, Jan 2013.

[179] Sae-Young Chung, G. D. Forney, T. J. Richardson, and R. Urbanke. On the Design of Low-Density Parity-Check Codes Within 0.0045 dB of the Shannon Limit. IEEE Communications Letters, 5(2):58–60, 2001.

[180] Amir Said. Introduction to Arithmetic Coding-Theory and Practice. Hewlett Packard Laboratories Report, pages 1057–7149, 2004.

[181] A. A. M. Saleh, A. Rustako, and R. Roman. Distributed Antennas for Indoor Radio Communications. IEEE Transactions on Communications, 35(12):1245–1251, 1987.

[182] D. Samardzija, J. Pastalan, M. MacDonald, S. Walker, and R. Valenzuela. Compressed Transport of Baseband Signals in Radio Access Networks. IEEE Transactions on Wireless Communications, 11(9):3216–3225, 2012.

[183] A. Sanderovich, S. Shamai, Y. Steinberg, and G. Kramer. Communication via Decentralized Processing. IEEE Transactions on Information Theory, 54(7):3008–3023, 2008.

[184] M. Sarajlić, L. Liu, and O. Edfors. When Are Low Resolution ADCs Energy Efficient in Massive MIMO? IEEE Access, 5:14837–14853, 2017.

[185] H. Sari, G. Karam, and I. Jeanclaude. Transmission Techniques for Digital Terrestrial TV Broadcasting. IEEE Communications Magazine, 33(2):100–109, 1995.

[186] A. Saul. Generalized Active Constellation Extension for Peak Reduction in OFDM Systems. In IEEE International Conference on Communications, 2005. ICC 2005. 2005, volume 3, pages 1974–1979 Vol. 3, 2005.

[187] Louis L Scharf. Statistical Signal Processing, volume 98. Addison-Wesley Reading, MA, 1991.

[188] I. D. Schizas, G. B. Giannakis, and Z. Luo. Distributed Estimation Using Reduced-Dimensionality Sensor Observations. IEEE Transactions on Signal Processing, 55(8):4284–4299, 2007.

[189] M Series. IMT Vision–Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond. Recommendation ITU, 2083, 2015.

[190] Seung Hee Han and Jae Hong Lee. An Overview of Peak-to-Average Power Ratio Reduction Techniques for Multicarrier Transmission. IEEE Wireless Communications, 12(2):56–65, 2005.

[191] Claude E. Shannon. A Mathematical Theory of Communication. Bell System Technical Journal, 27(3):379–423, 1948.

[192] Claude E. Shannon. Communication in the Presence of Noise. Proceedings of the IRE, 37(1):10–21, 1949.

[193] C. Shepard, R. Doost-Mohammady, J. Ding, R. E. Guerra, and L. Zhong. ArgosNet: A Multi-Cell Many-Antenna MU-MIMO Platform. In 2018 52nd Asilomar Conference on Signals, Systems, and Computers, pages 2237–2241, 2018.

[194] Y. Shi, J. Zhang, and K. B. Letaief. Group Sparse Beamforming for Green Cloud-RAN. IEEE Transactions on Wireless Communications, 13(5):2809–2823, 2014.

[195] Shidong Zhou, Ming Zhao, Xibin Xu, Jing Wang, and Yan Yao. Distributed Wireless Communication System: A New Architecture for Future Public Wireless Access. IEEE Communications Magazine, 41(3):108–113, 2003.

[196] I. Shomorony and A. S. Avestimehr. Worst-Case Additive Noise in Wireless Networks. IEEE Transactions on Information Theory, 59(6):3833–3847, June 2013.

[197] O. Simeone, U. Spagnolini, Y. Bar-Ness, and S. H. Strogatz. Distributed Synchronization in Wireless Networks. IEEE Signal Processing Magazine, 25(5):81–97, 2008.

[198] B. Sklar. A Primer on Turbo Code Concepts. IEEE Communications Magazine, 35(12):94–102, 1997.

[199] D. Slepian and J. Wolf. Noiseless Coding of Correlated Information Sources. IEEE Transactions on Information Theory, 19(4):471–480, 1973.

[200] Q. H. Spencer, C. B. Peel, A. L. Swindlehurst, and M. Haardt. An Introduction to the Multi-User MIMO Downlink. IEEE Communications Magazine, 42(10):60–67, Oct 2004.

[201] G. L. Stuber, J. R. Barry, S. W. McLaughlin, Ye Li, M. A. Ingram, and T. G. Pratt. Broadband MIMO-OFDM Wireless Communications. Proceedings of the IEEE, 92(2):271–294, 2004.

[202] C. Studer and G. Durisi. Quantized Massive MU-MIMO-OFDM Uplink. IEEE Transactions on Communications, 64(6):2387–2399, 2016.

[203] C. Studer and E. G. Larsson. PAR-Aware Large-Scale Multi-User MIMO-OFDM Downlink. IEEE Journal on Selected Areas in Communications, 31(2):303–313, 2013.

[204] S. Sun, T. S. Rappaport, T. A. Thomas, A. Ghosh, H. C. Nguyen, I. Z. Kovács, I. Rodriguez, O. Koymen, and A. Partyka. Investigation of Prediction Accuracy, Sensitivity, and Parameter Stability of Large-Scale Propagation Path Loss Models for 5G Wireless

Communications. IEEE Transactions on Vehicular Technology, 65(5):2843–2860, May 2016.

[205] S. Suyama, H. Adachi, H. Suzuki, and K. Fukawa. PAPR Reduction Methods for Eigenmode MIMO-OFDM Transmission. In VTC Spring 2009 - IEEE 69th Vehicular Technology Conference, pages 1–5, 2009.

[206] M. Suzuki, Y. Kishiyama, and K. Higuchi. Combination of Beamforming With Per-Antenna Power Constraint and Adaptive PAPR Reduction Method Using Null Space in MIMO Channel for Multiuser Massive MIMO-OFDM Transmission. In 2018 21st International Symposium on Wireless Personal Multimedia Communications (WPMC), pages 134–139, 2018.

[207] Taesang Yoo and A. Goldsmith. On the Optimality of Multiantenna Broadcast Scheduling Using Zero-Forcing Beamforming. IEEE Journal on Selected Areas in Communications, 24(3):528–541, 2006.

[208] Emre Telatar. Capacity of Multi-Antenna Gaussian Channels. European Transactions on Telecommunications, 10(6):585–595, 1999.

[209] Inc. The MathWorks. comm.SphereDecoder.

[210] K. T. Truong and R. W. Heath. The Viability of Distributed Antennas for Massive MIMO Systems. In 2013 Asilomar Conference on Signals, Systems and Computers, pages 1318–1323, 2013.

[211] D. N. C. Tse, P. Viswanath, and Lizhong Zheng. Diversity-Multiplexing Tradeoff in Multiple-Access Channels. IEEE Transactions on Information Theory, 50(9):1859–1874, 2004.

[212] David Tse and Pramod Viswanath. Fundamentals of Wireless Communication. Cambridge University Press, 2005.

[213] M. Tuchler, R. Koetter, and A. C. Singer. Turbo Equalization: Principles and New Results. IEEE Transactions on Communications, 50(5):754–767, 2002.

[214] G. L. Turin, F. D. Clapp, T. L. Johnston, S. B. Fine, and D. Lavry. A Statistical Model of Urban Multipath Propagation. IEEE Transactions on Vehicular Technology, 21(1):1–9, 1972.

[215] R. Van Nee, A. Van Zelst, and G. Awater. Maximum Likelihood Decoding in a Space Division Multiplexing System. In VTC2000-Spring. 2000 IEEE 51st Vehicular Technology Conference Proceedings (Cat. No.00CH37026), volume 1, pages 6–10 vol.1, 2000.

[216] M. K. Varanasi and T. Guess. Optimum Decision Feedback Multiuser Equalization With Successive Decoding Achieves the Total Capacity of the Gaussian Multiple-Access Channel. In Conference Record of the Thirty-First Asilomar Conference on Signals, Systems and Computers (Cat. No.97CB36136), volume 2, pages 1405–1409 vol.2, 1997.

[217] J. Vieira, S. Malkowsky, K. Nieman, Z. Miers, N. Kundargi, L. Liu, I. Wong, V. Öwall, O. Edfors, and F. Tufvesson. A Flexible 100-Antenna Testbed for Massive MIMO. In 2014 IEEE Globecom Workshops (GC Wkshps), pages 287–293, 2014.

[218] P. Viswanath and D. N. C. Tse. Sum Capacity of the Vector Gaussian Broadcast Channel and Uplink-Downlink Duality. IEEE Transactions on Information Theory, 49(8):1912–1921, Aug 2003.

[219] A. Wakeel and W. Henkel. Least-Squares Iterative PAR Reduction for Point-to-Point Large-Scale MIMO-OFDM Systems. In 2014 IEEE International Conference on Communications (ICC), pages 4638–4643, 2014.

[220] A. Wiesel, Y. C. Eldar, and S. Shamai. Zero-Forcing Precoding and Generalized Inverses. IEEE Transactions on Signal Processing, 56(9):4409–4418, Sep. 2008.

[221] F. Wiffen, M. Z. Bocus, A. Doufexi, and W. H. Chin. MF-Based Dimension Reduction Signal Compression for Fronthaul-Constrained Distributed MIMO C-RAN. In 2020 IEEE Wireless Communications and Networking Conference (WCNC), pages 1–8, 2020.

[222] F. Wiffen, M. Z. Bocus, A. Doufexi, and A. Nix. Phase-Only OFDM Communication for Downlink Massive MIMO Systems. In 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), pages 1–5, 2018.

[223] F. Wiffen, M. Z. Bocus, A. Doufexi, and A. Nix. Distributed MIMO Uplink Capacity Under Transform Coding Fronthaul Compression. In ICC 2019 - 2019 IEEE International Conference on Communications (ICC), pages 1–6, 2019.

[224] F. Wiffen, L. Sayer, M. Z. Bocus, A. Doufexi, and A. Nix. Comparison of OTFS and OFDM in Ray Launched Sub-6 GHz and mmWave Line-of-Sight Mobility Channels. In 2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), pages 73–79, 2018.

[225] C. Windpassinger, R. F. H. Fischer, T. Vencel, and J. B. Huber. Precoding in Multiantenna and Multiuser Communications. IEEE Transactions on Wireless Communications, 3(4):1305–1316, July 2004.

[226] A. Winkelbauer, S. Farthofer, and G. Matz. The Rate-Information Trade-Off for Gaussian Vector Channels. In 2014 IEEE International Symposium on Information Theory, pages 2849–2853, 2014.

[227] A. Winkelbauer and G. Matz. Rate-Information-Optimal Gaussian Channel Output Compression. In 2014 48th Annual Conference on Information Sciences and Systems (CISS), pages 1–5, 2014.

[228] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Valenzuela. V-BLAST: An Architecture for Realizing Very High Data Rates Over the Rich-Scattering Wireless Channel. In 1998 URSI International Symposium on Signals, Systems, and Electronics. Conference Proceedings (Cat. No.98EX167), pages 295–300, Oct 1998.

[229] M. Wu, B. Yin, G. Wang, C. Dick, J. R. Cavallaro, and C. Studer. Large-Scale MIMO Detection for 3GPP LTE: Algorithms and FPGA Implementations. IEEE Journal of Selected Topics in Signal Processing, 8(5):916–929, 2014.

[230] Q. Wu, G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober. An Overview of Sustainable Green 5G Networks. IEEE Wireless Communications, 24(4):72–80, 2017.

[231] G. Wunder, R. F. H. Fischer, H. Boche, S. Litsyn, and J. No. The PAPR Problem in OFDM Transmission: New Directions for a Long-Lasting Problem. IEEE Signal Processing Magazine, 30(6):130–144, 2013.

[232] A. Wyner and J. Ziv. The Rate-Distortion Function for Source Coding With Side Information at the Decoder. IEEE Transactions on Information Theory, 22(1):1–10, 1976.

[233] X. Xu, X. Rao, and V. K. N. Lau. Active User Detection and Channel Estimation in Uplink CRAN Systems. In 2015 IEEE International Conference on Communications (ICC), pages 2727–2732, 2015.

[234] B. Yang, G. Mao, M. Ding, X. Ge, and X. Tao. Dense Small Cell Networks: From Noise-Limited to Dense Interference-Limited. IEEE Transactions on Vehicular Technology, 67(5):4262–4277, 2018.

[235] H. Yang and T. L. Marzetta. Performance of Conjugate and Zero-Forcing Beamforming in Large-Scale Antenna Systems. IEEE Journal on Selected Areas in Communications, 31(2):172–179, 2013.

[236] H. Yang and T. L. Marzetta. A Macro Cellular Wireless Network With Uniformly High User Throughputs. In 2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall), pages 1–5, 2014.

[237] H. Yang and T. L. Marzetta. Massive MIMO With Max-Min Power Control in Line-of-Sight Propagation Environment. IEEE Transactions on Communications, 65(11):4685–4693, 2017.

[238] S. Yang and L. Hanzo. Fifty Years of MIMO Detection: The Road to Large-Scale MIMOs. IEEE Communications Surveys Tutorials, 17(4):1941–1988, 2015.

[239] M. Yao, M. Carrick, M. M. Sohul, V. Marojevic, C. D. Patterson, and J. H. Reed. Semidefinite Relaxation-Based PAPR-Aware Precoding for Massive MIMO-OFDM Systems. IEEE Transactions on Vehicular Technology, 68(3):2229–2243, 2019.

[240] Yi Jiang, Xiayu Zheng, and Jian Li. Asymptotic Performance Analysis of V-BLAST. In GLOBECOM '05. IEEE Global Telecommunications Conference, 2005., volume 6, pages 5 Pp.–3886, 2005.

[241] C. Yu and M. Yen. Area-Efficient 128- to 2048/1536-Point Pipeline FFT Processor for LTE and Mobile WiMAX Systems. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 23(9):1793–1800, 2015.

[242] W. Yu and T. Lan. Transmitter Optimization for the Multi-Antenna Downlink With Per-Antenna Power Constraints. IEEE Transactions on Signal Processing, 55(6):2646–2660, 2007.

[243] R. Zayani, H. Shaiek, and D. Roviras. PAPR-Aware Massive MIMO-OFDM Downlink. IEEE Access, 7:25474–25484, 2019.

[244] J. Zhang, L. Dai, X. Zhang, E. Björnson, and Z. Wang. Achievable Rate of Rician Large-Scale MIMO Channels With Transceiver Hardware Impairments. IEEE Transactions on Vehicular Technology, 65(10):8800–8806, 2016.

[245] M. Zhao and A. Burr. Compression and Recovery Scheme for Cell-Free Cloud Radio Access Network. In 2019 IEEE Wireless Communications and Networking Conference (WCNC), pages 1–6, 2019.

[246] Zhendao Wang and G. B. Giannakis. Wireless Multicarrier Communications. IEEE Signal Processing Magazine, 17(3):29–48, 2000.

[247] Y. Zhou, Y. Xu, J. Chen, and W. Yu. Optimality of Gaussian Fronthaul Compression for Uplink MIMO Cloud Radio Access Networks. In 2015 IEEE International Symposium on Information Theory (ISIT), pages 2241–2245, 2015.

[248] Y. Zhou and W. Yu. Optimized Beamforming and Backhaul Compression for Uplink MIMO Cloud Radio Access Networks. In 2014 IEEE Globecom Workshops (GC Wkshps), pages 1493–1498, 2014.

[249] Y. Zhou and W. Yu. Fronthaul Compression and Transmit Beamforming Optimization for Multi-Antenna Uplink C-RAN. IEEE Transactions on Signal Processing, 64(16):4138–4151, 2016.